L'utilisation du Natural Langage Processing comme outil d'aide à l'exploitation de données hospitalières

CUNIN Marie-Pierre[1], BODART Gwennaëlle[3], THYS Marie[2], GANGOLF Marjorie[2], KOLH Philippe[4], JACQUES Jessica [2]

[1] CHU Liège, Service des Informations Médico-Economiques, Avenue de l'hôpital 4000 Liège, 0032(0)4/3425582, mpcunin@chuliege.be
[2] CHU Liège, Service des Informations Médico-Economiques, Avenue de l'hôpital 4000 Liège,
[3] CHU Liège, Institut de Cancérologie, Avenue de l'hôpital 4000 Liège,
[4] CHU Liège, Gestion du Système d'Informations, Avenue de l'hôpital 4000 Liège

Résumé.

Le CHU de Liège dispose d'un Dossier Médical Informatisé dans lequel coexistent des données structurées et non structurées. Depuis plus de 10 ans, il a aussi développé un large entrepôt de données clinques compilant l'ensemble de l'information clinique et paraclinique de l'institution. De plus, une équipe d'analystes est devenue experte de l'exploitation des données, essentiellement des données structurées.
La question qui se pose actuellement est de savoir comment tirer et exploiter la richesse de toute l'information contenue dans les données non structurées du Dossier Patient Informatisé. Plus particulièrement, les nouvelles technologies de traitement du langage naturel (NLP Natural Langage Processing) apportent-elles une plus-value significative dans l'exploitation des données médicales ?

Mots clés: NLP, extraction de données, SNOMED.

1. Introduction

Le CHU de Liège est le seul hôpital universitaire en Wallonie associé à un cycle complet de la Faculté de médecine. Il s'agit un hôpital pluridisciplinaire, dont les 3 principales missions sont les activités cliniques, l'enseignement et la recherche. Celles-ci sont réparties sur 7 sites, dont 4 d'hospitalisation. Il dispose de 1.038 lits d'hospitalisation agréés dont 828 lits aigus parmi lesquels 49 lits de soins intensifs et compte près de 5 700 ETP salariés. Plus de 900 médecins, répartis entre une cinquantaine de services, y dispensent des soins spécialisés dans toutes les disciplines médicales.

Le Service des informations médico-économiques (SIME) est divisé en quatre secteurs distincts :

- Codage et Nomenclature
- Exploitation des Données
- Appui à la Recherche Clinique et Biostatistique
- Gestion des Dossiers Médicaux (archives médicales)

Plus spécifiquement, le secteur Exploitation des Données a pour mission de mettre à disposition les données médicales et analyser la situation médico-économique de l'hôpital. Parmi les demandes faites au secteur, on retrouve notamment l'identification de patients sur base d'une recherche de mots-clés à travers le dossier médical des patients ou sur la base de l'identification d'une pathologie.

Pour ce faire, différentes sources sont utilisées comme le dossier médical ou infirmier informatisé, les données administratives des patients, la facturation, les données de laboratoire ou d'imagerie médicale, le codage RHM[1], etc. Ces données sont regroupées dans un datawarehouse exploitable via SAS et SQL serveur notamment.

Dans ce cadre, extraire de la valeur informationnelle à partir des textes libres est une source d'information non négligeable. En effet, en clinique, les informations résident souvent dans les textes libres. L'analyse de ces textes présente donc un bénéfice supplémentaire en termes d'exploitation de données.

Néanmoins, une extraction de mots clés ne permet pas de traiter totalement certaines spécificités du langage telles que l'utilisation d'acronymes et/ou de synonymes, les fautes d'orthographe, les négations (lorsque la syntaxe de la phrase ne met pas directement en relation la négation et le mot recherché), le contexte familial, la conditionnalité, la temporalité, … .

Dès lors, au vu des diverses avancées réalisées dans le domaine du Clinical NLP (traitement du langage naturel), le CHU de Liège a souhaité disposer d'une solution permettant une couverture lexicale poussée ainsi qu'un traitement avancé du contexte linguistique et de la sémantique.

2. Contexte et objectif

Le souhait initial du CHU de Liège dans l'acquisition d'un outil de NLP était :

- D'indexer l'historique médical des patients, en particulier identifier les antécédents médicaux et chirurgicaux
- De codifier les concepts médicaux en SNOMED CT et, éventuellement, en ICD-10-BE
- De valider et améliorer l'extraction de données
- D'offrir des fonctionnalités de navigation et de valorisation des données, dans un cadre médical (orientation patient) ou dans un cadre de gestion opérationnelle des hôpitaux (orientation institution)
- D'injecter le flux de données structurées produites dans l'architecture informatique de l'hôpital

Il a cependant été décidé qu'avant toute intégration de l'outil tant au niveau du dossier patient que du codage RHM, un premier test de son efficacité et de sa plus-value serait réalisé dans le cadre de l'exploitation des données.

Le CHU a acquis en novembre 2018 un outil de NLP, géré par un fournisseur externe. Il est à noter que ce fournisseur travaillant avec d'autres structures hospitalières disposait déjà d'un modèle entrainé avec des données cliniques, validés par les médecins de ces hôpitaux. En amont du travail d'indexation de nos données, le fournisseur a néanmoins réalisé un travail de personnalisation visant à identifier les éventuelles carences du modèle standard et ainsi capter les composantes linguistiques propres au CHU de Liège. Pour ce faire, une annotation manuelle d'un corpus de documents (1250 documents représentatifs des textes qui devront être annotés et codés) a été réalisée. Cette annotation manuelle a été comparée à l'annotation automatique générée par l'outil standard. Des statistiques (telles que des erreurs de type I et de type II, précision, rappel et F-measure) ont été produites et ont permis un ré-entrainement du modèle. Le modèle est en outre régulièrement ré-entrainé, sur base des documents fournis par le CHU de Liège ou d'autres hôpitaux.

1 Résumé Hospitalier Minimum qui est utilisé pour répartir une partie du financement des hôpitaux par le biais du «budget des moyens financiers » (BMF).

Ce premier ré-entrainement de personnalisation ayant été effectué, le secteur Exploitation des Données a alors mis à disposition du fournisseur NLP des données historiques dans un premier temps (2018 à 2020), issues du dossier médical informatisé, à savoir :

- les courriers (lettres de consultation, rapports d'hospitalisation, certificats médicaux…)
- les protocoles opératoires
- les tours de salle

Ces documents ont été indexés et analysés par le fournisseur NLP. Les résultats, sous formes d'annotations en code SNOMED et ICD-10, ont ensuite été récupérés et stockés dans le DWH du CHU de Liège.

S'en sont suivies trois validations. Celles-ci ont concerné des thèmes différents et ont été réalisées conjointement par l'équipe du secteur avec l'aide d'un assistant en médecine et deux étudiantes en Master en Sciences de la Santé publique.

Plus particulièrement, ces validations ont consisté à comparer, pour trois thèmes d'extraction de données, les résultats générés par le NLP avec ceux générés par le secteur. Les thèmes choisis sont :

- Une recherche de mots-clés avec lien sémantique visant à identifier les patients atteints de la Maladie de Verneuil
- Deux recherches de pathologies chroniques: patients atteints de diabète et patients oncologiques.

3. Méthodologie

L'exercice de validation a porté sur les documents datés de 2018 envoyés au fournisseur NLP. Pour chacun des trois thèmes, le schéma de validation a été sensiblement le même, c'est-à-dire :

1. Extraction réalisée par le Secteur Exploitation des Données :
   1.1. recherche de mots clés dans le cas de la maladie de Verneuil,
   1.2. extraction multi-sources dans le cas du diabète (données de biologie, facturation dans la convention et anamnèse infirmière)
   1.3. extraction des patients catégorisés oncologiques au CHU de Liège selon une définition institutionnelle[2]. Ces patients sont identifiés pour une période de 5 années sur base de trois sources : les enregistrements cancers, le codage RHM et les traitements de radiothérapie, par ordre de priorité.

   Pour les trois thèmes, seuls les patients ayant eu un contact avec le CHU de Liège en 2018 ont été conservés.

2. Recherche des codes snomed associés aux pathologies visées et extraction des patients tagués sur base de ces codes dans la base de données issue du traitement NLP, avec exclusion des tags 'antécédents familiaux' et 'faux positif'.

2 Tout patient avec un nouveau diagnostic ou une récidive de cancer pris en charge dans l'institution (diagnostic et/ou traitement) est considéré comme patient oncologique pour une période de -2 mois à +5 ans autour de la date d'incidence/récidive (nommée fenêtre oncologique). Cette définition inclut toutes les tumeurs malignes invasives, les tumeurs malignes in situ, les tumeurs du système nerveux central (C70-71-72), les tumeurs neuroendocrines et les tumeurs incertaines urothéliales (C65-C68) et des ovaires (C56.9)

3. Quantification des différences en nombre de patients trouvés entre les deux méthodes, au global et éventuellement par sous-groupe (comme les types de cancer en oncologie)

4. Identification des matchs et des non-matchs de part et d'autre, comme indiqué dans la figure 2.

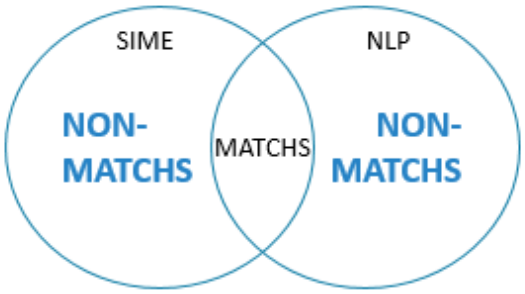

Figure 1. Sphères de validation

5. Revue de dossier de tous les non-matchs (dans le cas de la maladie de Verneuil) ou d'un échantillon (dans le cas du diabète et de l'oncologie) pour, d'une part identifier, les raisons des différences observées entre les deux méthodes et, d'autre part, valider que le lien fait entre le patient et la pathologie est correct (quelle que soit la méthode)

6. Classification et quantification de ces raisons sous forme de tableaux de fréquence.

4. Résultats

La comparaison entre les deux méthodes d'extraction a généré les résultats suivants pour chaque thème :

Maladie de Verneuil : sur le total des patients identifiés par les deux méthodes, 63% sont en commun. L'exploitation des données a généré plus de résultats que le NLP.

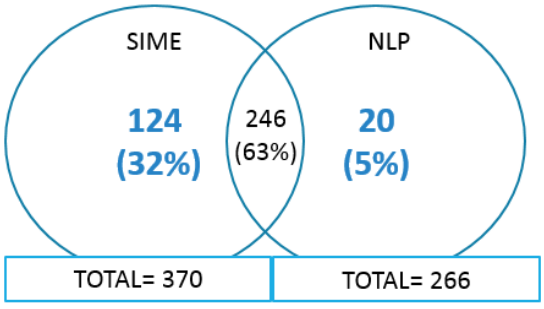

Figure 2. Résultats de la comparaison pour la maladie de Verneuil

Tous les non-matchs ont fait l'objet d'une revue de dossier.

Patients oncologiques : sur le total des patients identifiés par les deux méthodes, 36% sont en commun. Le NLP a généré 37% de patients de plus que l'exploitation des données. Or, la définition du patient oncologique

au CHU de Liège étant une définition institutionnelle validée par le département oncologique, elle constitue une « valeur étalon » pour la validation des résultats générés par le NLP.

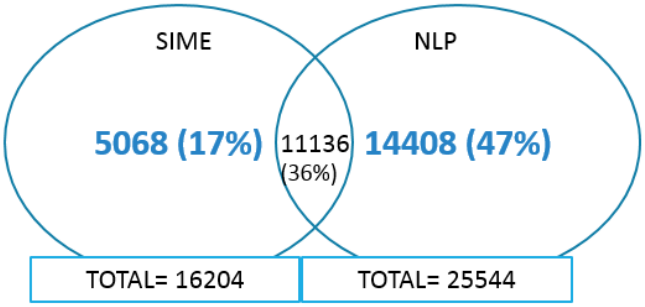

Figure 3. Résultats de la comparaison pour l'oncologie

La validation a consisté en la vérification des matchs et des non-matchs:

- Les patients repérés par le NLP mais pas par le CHU de Liège selon la définition institutionnelle : une première vérification a été réalisée sur base du diagnostic afin de valider que celui identifié par le NLP sont bien inclus dans la définition. 105 patients avaient des diagnostics exclus (ex: la plupart des tumeurs bénignes ou les carcinomes basocellulaires de la peau), soit 0.7% des cas non identifiés par le CHU. Pour les autres patients, 620 revues de dossiers ont été réalisées.

- Les patients repérés par l'Exploitation des Données et pas par le NLP: une première vérification quantitative des différences a été réalisée sur base de la fenêtre oncologique, de la source de données. Il est apparu de manière logique que plus l'évènement oncologique est éloigné de 2018, plus la proportion de cas non identifiés par le NLP est élevée. En effet, plus on s'éloigne du diagnostic, plus il y a de chance que le patient ne soit plus suivi pour son cancer. Il y a également les cancers diagnostiqués en janvier-février 2019, pour lesquels la fenêtre oncologique est ouverte 2 mois avant, c'est-à-dire en 2018. Il n'est donc pas surprenant qu'ils ne soient pas détectés par le NLP sur base de documents datant de 2018, puisque c'est avant leur diagnostic.

Année événement oncologique	N	%
2013	720	41,8%
2014	820	40,3%
2015	890	38,2%
2016	1045	35,2%
2017	878	21,2%
2018	821	14,1%
2019	333	45,5%
Total général	**5068**	**31,3%**

Tableau 1. Proportion de cas non identifiés par le NLP selon l'année de l'évènement oncologique

En outre, on observe que ce sont les cas identifiés via les sources RHM et la radiothérapie qui sont le plus souvent non retrouvés par le NLP, ce qui est logique également étant donné les documents auquel le fournisseur NLP a accès.

	Tous	
Source	**N**	**%**
Registre du cancer 1er	2945	28,8%

Registre du cancer Follow Up	490	16,1%
RHM	625	42,4%
Radiothérapie	1291	35,8%
Total général	**5068**	**31,3%**

Tableau 2. Proportion de cas non identifiés par le NLP selon la source

De plus, les types de cancer qui sont le moins souvent identifiés par le NLP sont les cancers neurologiques, gynécologiques et dermatologiques, avec 50% ou plus non identifiés, même pour les diagnostics de 2018 (cf tableau 3). Une revue de dossier a dès lors été réalisée sur 95 cas relevant de ces trois pathologies et a montré que dans près de 9 cas sur 10, la raison était extérieure au fonctionnement du modèle (cf tableau 4)

Catégorie de cancer	N	%
Neurologie	300	56,6%
Gynécologie	796	50,6%
Dermatologie	675	49,6%
Sénologie	1167	34,9%
Endocrinologie	111	28,3%
Urologie	767	27,7%
ORL et oncologie tête et cou	255	25,8%
Sarcome	33	25,2%
Digestive	367	18,3%
Pneumologie	268	18,3%
Hématologie	306	16,9%
Inconnu	113	16,7%

Tableau 3. Proportion de cas non identifiés par le NLP selon la catégorie de cancer

Raisons de non-identification	N	% des cas vérifiés
Code SNOMED non listé par le CHU	51	53,7%
Non identifiable par le NLP	33	34,7%
Terme non détecté par le NLP	6	6,3%
Erreur d'interprétation du NLP	4	4,2%
Patient pas onco (erreur CHU)	1	1,1%
Total général	**95**	**100,0%**

Tableau 4. Raisons de non-identification par le NLP pour les cancers neurologiques, gynécologiques et dermatologiques

Enfin, ce sont essentiellement les cancers non invasifs qui sont moins bien détectés par le NLP avec 50% ou plus non retrouvés.

Comportement	N	%
Bénin	221	55,39%
In situ	874	52,59%
Incertain	117	57,07%
Invasif	3931	27,52%
Total général	**5068**	**31,28%**

Tableau 5. Proportion de cas non identifiés par le NLP selon le comportement tumoral

Patients diabétiques : sur le total des patients identifiés par les deux méthodes, 29% sont en commun. Le NLP a généré 67% de patients de plus que l'exploitation des données.

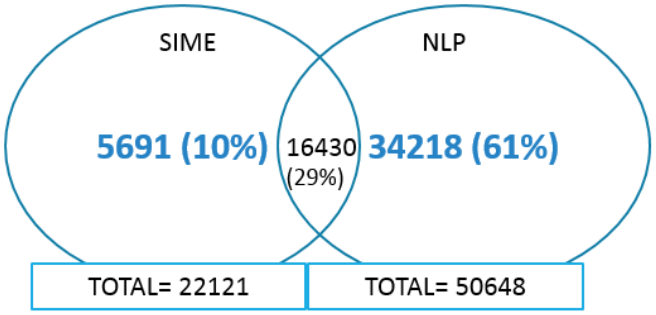

Figure 5. Résultats de la comparaison pour le diabète

Les non-matchs ont fait l'objet d'une revue de dossier :

- Les patients repérés par le NLP mais pas par l'Exploitation des Données : revue de dossier sur une sélection aléatoire de 100 patients.
- Les patients repérés par l'Exploitation des Données et pas par le NLP: revue de dossier sur une sélection aléatoire de 20 patients par source de données, soit 60 dossiers au total.

La revue de dossier des cas identifiés par le NLP mais pas par le CHU a fait émerger trois constats communs à tous les thèmes :

1. L''identification par le NLP des antécédents, personnels et familiaux (ces écueils ont été repérés dans respectivement 42% et 26% des revues de dossiers oncologiques). A noter qu'une mise à jour de cette validation avec une version plus récente de l'outil a montré une amélioration de 14% dans le repérage des antécédents personnels et 75% des antécédents familiaux.

2. Le repérage des faux positifs, qui altère la précision des résultats de l'indexation. Les faux-positifs représentent 19% des revues de dossiers des cas oncologiques et 20% des revues de dossier de la maladie de Verneuil.
   Cela correspond par exemple à la prise en compte lexicale de formulations simples telles que « possible » ou « suspicion de » ou plus complexes comme « image évoquant », « diagnostic de xxx écarté » ou «mutation génétique prédisposant à ».
   Des faux-positifs se retrouvent également lorsque la pathologie est reprise dans une référence de la littérature comme cela peut être fait dans des rapports de biologie génétique (« ce gène est décrit dans la maladie xxx »), ou dans des normes d'interprétation (par exemple, le mot « diabète » présent dans les normes d'analyses de laboratoire, et qui a généré 63% des cas repérés par le NLP et pas par l'Exploitation des Données). A noter que la mise à jour de cette validation avec une version plus récente de l'outil a montré une amélioration de 50% dans le repérage des faux-positifs (hors « diabète »).

3. Le lien sémantique entre le mot repéré par le NLP et le code SNOMED associé. Cet écueil peut prendre plusieurs formes.
   Dans le cas de la maladie de Verneuil, le dictionnaire sémantique utilisé par le programme NLP associe au code SNOMED de l'« Hidradenitis suppurativa (disorder) » les termes « Maladie de Verneuil », « Acné inversé » et « Hidrosadénite suppurée ». Ce dictionnaire est moins complet que la liste de mots-clés utilisée

par l'Exploitation des Données, liste qui a été fournie par le clinicien demandeur de l'extraction et qui inclut des termes %Verneuil%, %Hidrosadénite supp%, %acné inversé%, %Hidradénite supp% et % HS %. La recherche par mots-clés a ainsi permis d'identifier correctement 17 patients de plus que le NLP.
Dans le cas de l'oncologie, il est apparu dans 12% des revues de dossiers que le mot est incorrectement associé au code SNOMED. C'est le cas des acronymes (par exemple, PEM pour potentiel évoqué moteur qui est associé au code SNOMED « Malignant plasma cell neoplasm » ou encore l'acronyme mm utilisé pour millimètres associé au code SNOMED « Multiple myeloma (disorder) »). Un autre exemple, plus déroutant mais qui a généré l'indexation de 4536 patients, est l'association du mot « pomme » au code SNOMED « Pancreatic polypeptidoma (disorder) ». A noter que la mise à jour de cette validation avec une version plus récente de l'outil a montré une amélioration de 70% dans l'association lexicale.

Plus spécifique à l'oncologie, la validation a également montré la limitation de l'outil de NLP à repérer des concepts dans le cas où les informations sont dispersées dans le texte libre. En effet, la terminologie en oncologie est complexe et comprend une triple composante localisation-morphologie-comportement. Les codes Snomed associent les 3 aspects dans un même code. Or dans les courriers, les 3 composantes sont rarement explicitement indiquées ou ne se trouvent pas nécessairement dans la même phrase. Par exemple, dans un texte tel que « patient avec une volumineuse lésion dans le poumon droit […] Une biopsie a été réalisée […] L'anapathologie révèle un carcinome épidermoïde peu différencié » où la localisation et le comportement se trouvent dans des phrases différentes, le NLP les repère séparément mais ne les associe pas et ne fait donc pas le lien avec le code Snomed approprié. En revanche, le terme « cancer du poumon » est parfaitement identifié. Il y a donc ici un paradoxe : plus la formulation est descriptive et précise, moins l'outil de NLP se révèle performant.

La revue de dossier des cas identifiés par l'Exploitation des Données mais pas par le NLP a fait émerger d'autres constats communs à tous les thèmes :

1. La sélection des codes Snomed, inhérente à l'identification de la pathologie. Dans le cas de pathologies peu communes ou de concepts très spécifiques (comme la maladie de Verneuil), la sélection du/des codes est sans ambiguïté et facilement exhaustive. Les résultats de l'indexation par le NLP sont dès lors performants. En revanche, dans le cas d'un concept large et complexe comme l'oncologie, la sélection a priori des codes SNOMED est plus compliquée. Dans notre cas, la requête utilisée pour générer les codes SNOMED oncologiques correspondant à la définition du CHU de Liège a retourné plus de 5 000 codes. Vérifier l'exactitude et l'exhaustivité d'une telle liste est un travail complexe et chronophage, mais pour autant indispensable pour s'assurer de la performance de l'indexation NLP. Dans le cas de l'oncologie, l'absence de codes SNOMED explique 54% des revues de dossiers des cas repérés par l'Exploitation des Données et pas par le NLP (cf tableau 4).

2. La source de données. En effet, le fournisseur NLP ayant accès à un panel limité d'informations, il est logique que certaines informations puissent être disponibles ailleurs et dès lors non repérées par le NLP, comme cela a été évoqué plus haut. Néanmoins, même lorsque les sources sont a priori les mêmes (le dossier médical informatisé), il peut y avoir des différences. Il s'agit par exemple de cas où le courrier n'a pas été généré suite à une consultation. En effet, dans le dossier médical informatisé, l'événement dans lequel le clinicien encode les informations et le courrier qui en est généré sont deux objets différents. L'Exploitation des Donnés extrait les données de l'événement qui est structuré en champs alors que le NLP exploite le texte du courrier.

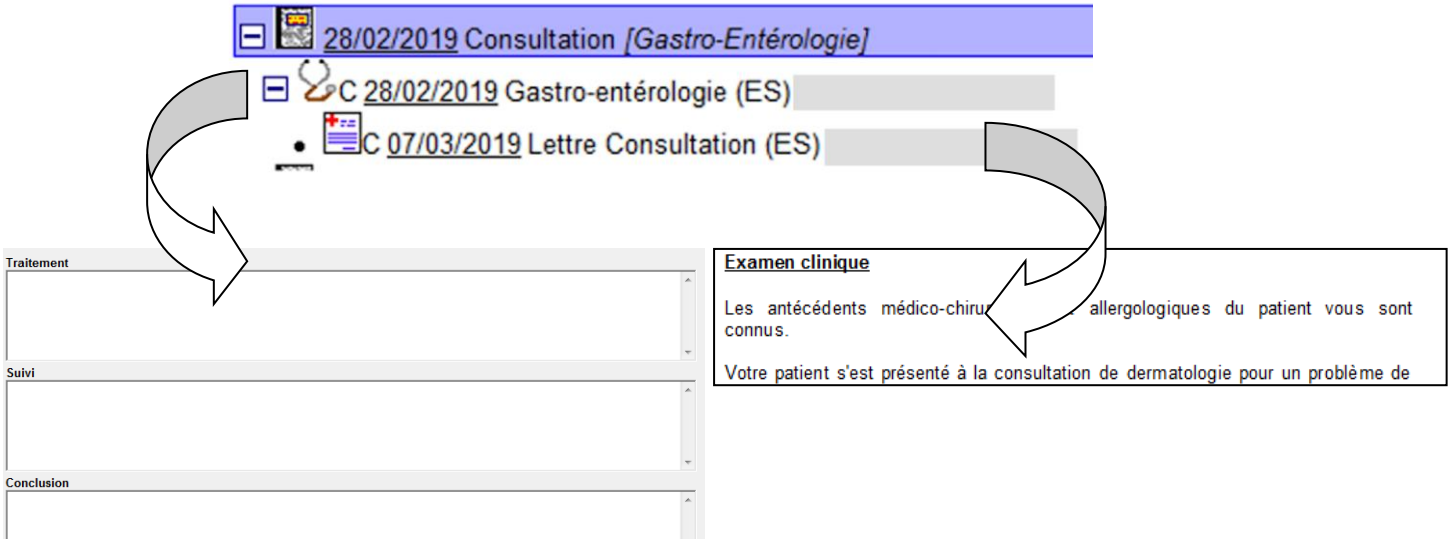

Figure 3. Evénement et courrier dans le DMI

A l'inverse, les courriers peuvent être retravaillés manuellement et donc certaines informations peuvent s'y trouver alors qu'elles ne sont pas dans l'événement lui-même. Dès lors, l'exploitation des données peut avoir accès à des informations dont le NLP ne dispose pas et inversement.
La différence de source représente 41% des non-matchs dans le cadre de la recherche de la maladie de Verneuil et 35% des revues de dossiers des cas oncologiques (cf tableau 4).

5. Conclusions

Dans le cadre d'une recherche de mots clés, et dans le cas de concepts très spécifiques, le NLP testé au CHU de Liège s'est révélé précis, tant que la panoplie des termes utilisés par les utilisateurs correspond à la liste des termes associés au code Snomed. Dans le cadre de l'Exploitation des Données, le NLP est utile et apporte des éléments supplémentaires mais relativement limités par rapport à la méthode actuelle par recherche de mots-clés.

En revanche, l'outil de NLP présente moins de précision dans le cas de concepts larges et complexes comme l'oncologie. Cela est en partie dû à l'outil et à son niveau d'entrainement pour l'interprétation lexicale et le repérage de concepts complexes dans un langage non structuré, mais pas seulement puisque la sélection correcte et exhaustive des codes SNOMED de la pathologie recherchée et la pertinence des documents utilisés pour l'indexation représentent également des limitations potentielles.

Il apparaît donc clairement la nécessité d'améliorer les process/outils de part et d'autre pour optimiser l'outil de NLP. Des pistes d'amélioration ont été proposées au fournisseur en ce sens, telles qu'élargir le périmètre exploité en incluant d'autres documents comme les compte-rendus d'imagerie ou d'anapathologie, sélectionner les informations pertinentes au sein des événements (les conclusions par exemple) pour éviter le « bruit », que l'outil fournisse par patient un indice de confiance dans l'association avec la pathologie. L'outil étant en constante progression, l'objectif pourrait également être de monitorer au fil des nouvelles versions certaines des constations faites lors de la validation et de sa mise à jour.

L'utilisation d'un outil NLP tel que celui que le CHU de Liège a testé, dans sa maturité actuelle, va dès lors dépendre du contexte de la demande (besoin ou non d'exhaustivité par exemple). En effet, les résultats

générés peuvent y répondre de manière plus ou moins pertinente et performante en fonction du degré de précision ou de sensibilité souhaité.

Les enjeux du NLP dans le cadre de l'exploitation des données sont différents de ceux qui visent à utiliser le NLP dans le contexte opérationnel du Dossier patient Informatisé. En effet, avec l'évolution des DPI et leur remplacement progressif, nombre d'hôpitaux cherchent à alimenter un nouveau DPI à partir de techniques de NLP. Dans ce cas, l'exhaustivité et la précision des données cliniques sont cruciales. Dans le contexte de l'exploitation des données, le NLP doit prouver son efficacité en limitant au maximum les faux-positifs. L'exhaustivité est peut-être moins privilégiée. A ce titre, le NLP est un bon outil pour alimenter un entrepôt de données cliniques.