# Enrichissement sémantique associé à la détection de la négation et des antécédents familiaux dans un entrepôt de données hospitalier

## *Semantic enrichment based on detection of negation and history family in a French hospital data warehouse*

**Nicolas Garcelon[1, 2], Rémi Salomon[3], Anita Burgun[2]**

*[1]Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France*
*[2]INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Paris*
*[3]Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique - Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité -*

**Résumé**

(Bakground) Nous avons développé dans des travaux précédents un entrepôt de données biomédicales permettant de fouiller les comptes rendus par un moteur de recherche plein texte. Mais l'absence de prise en compte de la négation et des antécédents familiaux engendraient du bruit. (Methods) Nous avons développé un algorithme permettant de détecter en français la notion de négation dans les comptes rendus médicaux, ainsi que le contexte d'antécédents familiaux. L'enrichissement sémantique tient compte de ces nouvelles informations sémantiques. Nous intégrons dans le schéma de l'entrepôt de données ainsi que dans l'algorithme de fouille de données ces nouvelles informations afin d'améliorer les performances du moteur de recherche. (Results) L'intégration de la détection de la négation et des antécédents familiaux améliore nettement la précision du moteur de recherche malgré une légère diminution du rappel. (Conclusion) Cet enrichissement syntaxique et sémantique a été intégré dans l'entrepôt de données avec succès.

***Abstract***

*(Background) We developed in previous work a biomedical data warehouse for mining clinical reports by full text search engine. But the lack of consideration of negation and family history generate noise. (Methods) We developed an algorithm to detect the French notion of negation in medical reports and he context of family history. The semantic enrichment reflects this new semantic information. We integrate in the data warehouse schema and in data mining algorithm this new information to improve the performance of the search engine. (Results) the integration of the detection of negation and family history significantly improves the precision of the search engine despite a slight decrease in recall. (Conclusion) The syntactic and semantic enrichment has been successfully integrated into the data warehouse.*

**Mots-clés :** Entrepôt de données ; fouille de données ; Traitement automatique du langage

***Keywords:*** *Data warehouse; data mining; Natural Language Processing*

# 1 Introduction

Lors de travaux précédents, nous avons développé un entrepôt de données essentiellement orienté sur des comptes rendus plein texte [1]. Il s'agissait de proposer un outil simple permettant de rapidement fouiller les comptes rendus médicaux des patients. Dans l'onglet résultat du moteur de recherche, l'utilisateur peut facilement vérifier si le patient est vrai positif par l'affichage de la section du document contenant les termes recherchés. Nous avons ensuite développé un module d'enrichissement sémantique [2] utilisant le Metathesaurus de l'UMLS [3]. L'utilisateur choisit s'il veut utiliser l'enrichissement sémantique en fonction de la sensibilité et spécificité souhaitée. Le moteur de recherche exploite les relations de synonymie et de subsomption du Metathesaurus pour retrouver le plus grand nombre de patients dont les documents contiennent les termes recherchés, ce qui a pour effet d'augmenter la sensibilité aux dépends d'une moins bonne spécificité. Le bruit obtenu est lié non seulement aux phrases négatives (« absence de », « pas de » etc.) mais aussi au contexte d'antécédents familiaux dans lequel l'expression recherchée est située (« la sœur du patient est atteinte de »). Ce constat réalisé nous proposons une méthode pour détecter le contexte d'antécédents familiaux, ainsi que les sous phrases négatives afin de les exclure de la recherche plein texte. Ces méthodes sont particulièrement nécessaires pour traiter les comptes rendus produits dans les services prenant en charge des patients atteints de maladies génétiques héréditaires, comme l'Hôpital Necker à Paris et l'Institut Imagine (www.institutimagine.org/). Il s'agit aussi de compléter l'enrichissement sémantique par l'ajout de ce niveau de certitude et du contexte dans lequel le concept est retrouvé.

# 2 État de l'art

La plupart des travaux de Traitement Automatique du Langage concernent la langue anglaise. En 2001, Chapman et al. développent NegEx [4]. Il s'agit d'un algorithme qui se veut simple et ouvert. L'auteur associe une extraction sémantique à partir des concepts UMLS et une détection par expression régulière d'une forme négative avant ou après le concept UMLS extrait. Si la phrase ne contient pas de concept UMLS, la phrase n'est pas traitée. Goryachev et al. [5] ont évalué en 2006 4 méthodes de détection de la négation. Il a notamment comparé la méthode NegEx et la méthode NegExpander développée par Aronow et al. [6]. Tandis que NegEx utilise l'intégralité de la phrase, NegExpander se base sur l'analyse de syntagmes nominaux en découpant le texte suivant certains critères prédéfinis (ponctuation, « and », « or »). La détection de la négation est réalisée indépendamment de l'extraction de concepts UMLS. Les 2 autres méthodes décrites dans cet article se basent sur un système d'apprentissage automatique (Weka) sur un corpus de plus de 1700 résumés de sortie pour lesquels les 8 000 termes extraits ont été manuellement annotés par un expert comme un fait, une négation ou une éventualité. Ces 2 méthodes divergent ensuite par l'algorithme de classification : la classification bayésienne naïve et le Vector Space Model. Il conclut que la méthode NegEx reste la plus efficace. L. Deléger & Grouin ont proposé d'adapter NegEx à la langue française [7]. L'adaptation française de NegEx donne un rappel et une précision corrects. Toutefois, l'auteur fait remarquer que leur méthode ne fonctionne pas lorsque la phrase est trop longue et contient plusieurs informations négatives et affirmatives.

La méthode que nous développerons ici s'approche davantage de NegExpander adapté au français. De plus, notre méthode doit fonctionner indépendamment d'un enrichissement sémantique. C'est à dire que l'utilisateur pourra rechercher un terme dans le plein texte sans nécessairement qu'il y ait un concept UMLS retrouvé.

En ce qui concerne l'extraction automatique des antécédents familiaux dans les comptes rendus

médicaux, Friedlin & McDonald [8] ont utilisé l'outil REX (REgenstrief data eXtraction tool) afin d'associer 12 maladies au contexte d'antécédents familiaux avec une précision de 97% et un rappel de 96%. Goryachev et al. [9] ont proposé une méthode afin de déterminer le contexte d'antécédents familiaux des concepts UMLS extraits. Cette méthode consiste à découper le compte rendu en sections (« antécédents », « antécédents familiaux », etc.), puis chaque section est découpée en phrase puis en groupe nominal. Enfin il détermine par des règles de décision si la maladie détectée dans le groupe nominale est associée à un parent du patient ou au patient lui même avec une sensibilité de 97,2% et une spécificité de 99,7%. Lewis et al. [10] quant à lui prend l'intégralité du texte sans passer par un découpage en section. Il détecte les chaines de dépendances entre la relation de parenté et une maladie grâce au Stanford NLP Parser. Cela lui permet d'extraire des extraits de phrase contenant le lien entre un parent du patient et une maladie avec une précision de 61% et un rappel de 51%.

Notre méthode de détection des antécédents familiaux s'inspire davantage de la méthode NegEx utilisée pour la détection de la négation, car elle est beaucoup plus simple à mettre en œuvre et très facilement évolutive.

# 3 Matériel et méthodes

## 3.1 Prétraitement des textes

Il est indispensable d'effectuer un « nettoyage » des documents avant l'extraction des contextes et de la négation. En effet, les textes sont issus d'une conversion automatique de documents Word en format txt, ou sont issus de documents HTML. Une première étape consiste donc à supprimer les balises HTML, et surtout à réajuster les sauts de ligne. Suivant la version de Word, la conversion au format texte induit des sauts de ligne au milieu des phrases, ce qui peut modifier l'interprétation de l'algorithme. Il s'agit donc de déterminer si un saut de ligne doit être remplacé par un espace ou par un point de ponctuation. Et inversement, certains documents contiennent des pseudos tableaux ou des listes sur deux colonnes. La distinction entre deux propositions ou syntagmes nominaux n'est donc plus représentée par un saut de ligne ou par une ponctuation mais par plusieurs espaces. Par ailleurs, une phrase peut contenir des doubles espaces liés à une erreur de saisie. Nous avons donc développé quelques règles sous forme d'expression régulière qui prend en compte ces variabilités afin de recréer la phrase sans saut de ligne ou de forcer l'ajout d'un point séparant deux phrases distinctes. Cette étape est essentielle pour la suite des traitements.

## 3.2 Détection du contexte familial

Nous avons établi une liste d'expressions décrivant un lien de parenté avec le patient : sœur, frère, cousin, tante, père, mère, grand parent, oncle, cousin, antécédent familial, neveu, nièce, antécédents familiaux, papa, maman. Cette liste initiale peut facilement être complétée par la suite pour la rendre plus exhaustive. Les expressions régulières tiennent compte des variations orthographiques des formes nominales : féminin, masculin, singulier, pluriel, présence ou erreur sur l'accentuation : `[^a-z]fr[éeèê]res?`

L'algorithme découpe le texte en phrases en utilisant les éléments de ponctuation et les sauts de ligne du texte. Puis pour chaque phrase il vérifie si elle contient une des expressions exprimant la notion de famille ou un lien de parenté. La phrase est alors extraite de manière à obtenir à la fin de cette étape deux textes correspondants aux deux contextes 'patient' et 'antécédents familiaux'.

Nous avons choisi d'attribuer l'antécédent familial à l'intégralité de la phrase plutôt qu'à une sous partie de la phrase. En effet, nous avons testé les deux méthodes, et il apparaît plus que l'auteur du

compte rendu décrive les antécédents familiaux d'un seul bloc sans évoquer les données cliniques du patient. En considérant les propositions, nous attribuions trop souvent des antécédents familiaux au patient lui même car le lien de parenté n'est citée qu'une fois dans la phrase. Ainsi afin de réduire le nombre de faux négatifs dans la détection des antécédents familiaux et le nombre de faux positifs dans la recherche de patients, nous avons décidé d'utiliser le découpage par phrase.

## 3.3 Détection de la négation

La détection de la négation ressemble davantage à la méthode NegExpand. Il existe une grande hétérogénéité stylistique entre les comptes rendus médicaux, suivant les auteurs et les services. En raison de cette complexité syntaxique des comptes rendus dans lesquels une phrase peut exprimer plusieurs informations (affirmatives et négatives), nous avons créé une liste de règles permettant de découper ces phrases en se basant sur les expressions suivantes : `mais ; pour ; qui ; entre ; car ; , [^d] ; sans ; lequel : laquelle ; hormis ; parce qu ; bien qu ;en dehors : malgré ; en raison de.`

Nous découpons après une virgule uniquement si elle n'est pas suivie de la lettre « d », cela afin de ne pas découper les listes qui suivent une négation : « le patient n'a pas de diabète, d'insuffisance cardiaque, de cholestérol ». Nous avons créé une liste d'expressions régulières permettant de détecter une tournure négative. Les règles de détection de la négation prennent en compte des formes variées de la négation : `pas [a-z]* de, absence de, oriente pas vers, exclure, non, aucun, absence, absent, negati, elimine, infirme` etc. A cela nous avons ajouté des règles pour prendre en compte les doubles négations (« pas d'exclure », « pas d'éliminer ») ou les expressions exprimant une forte affirmation par l'utilisation de la négation : « pas de doute », « pas de soucis », « pas de problème ». Nous avons aussi tenu compte de l'expression de la « normalité » surtout utilisée en génétique : « gène OPA1 normal » qui signifie « non muté ». Nous avons donc considéré le terme « normal » comme l'expression de la négation d'un symptôme, d'une mutation ou d'un dysfonctionnement.

Nous considérons le triplet : texte-contexte-certitude afin de décrire le contenu d'un document. Suivant la complexité du document il sera décrit par plusieurs triplets, avec un maximum de 6 triplets :

[texte] - texte intégral -1 : le texte intégral

[texte] - texte intégral -0 : les sous phrases négatives du texte intégral

[texte] - texte patient -1 : le texte patient

[texte] - texte patient -0 : les sous phrases négatives du texte patient

[texte] - antécédents familiaux -1 : les antécédents familiaux

[texte] - antécédents familiaux -0 : les sous phrases négatives des antécédents familiaux

## 3.4 Enrichissement sémantique

Nous avons extrait le sous-ensemble français du Metathesaurus UMLS. Tous les concepts, leurs synonymes et leurs déclinaisons sont listés dans une table puis triés par ordre décroissant de longueur de chaine de caractères. La présence de chaque terme ou expression est testée dans le texte pour chaque triplet texte-contexte-certitude.

## 3.5 Intégration à l'entrepôt de données

L'entrepôt de données a été développé sous Oracle, en langage PHP, HTML, jQuery. Le schéma relationnel d'origine de la base de données a été modifié afin de pouvoir stocker les parties de texte reliées au contexte et au niveau de certitude et aux concepts éventuellement extraits du texte (**Erreur ! Source du renvoi introuvable.**). Nous avons donc (i) une table contenant le document originel à afficher, (ii) une table contenant le texte (indexé par Oracle text, et composé des phrases et sous phrases extraites par notre méthode TAL) en lien avec le contexte (patient, antécédents familiaux) et le niveau de certitude (0 ou 1) soit le triplet « texte-contexte-certitude », (iii) une table contenant les concepts extraits de chaque triplet. Oracle text indexe aussi la liste des synonymes de chaque concept UMLS afin de faciliter l'expansion sémantique d'une requête.

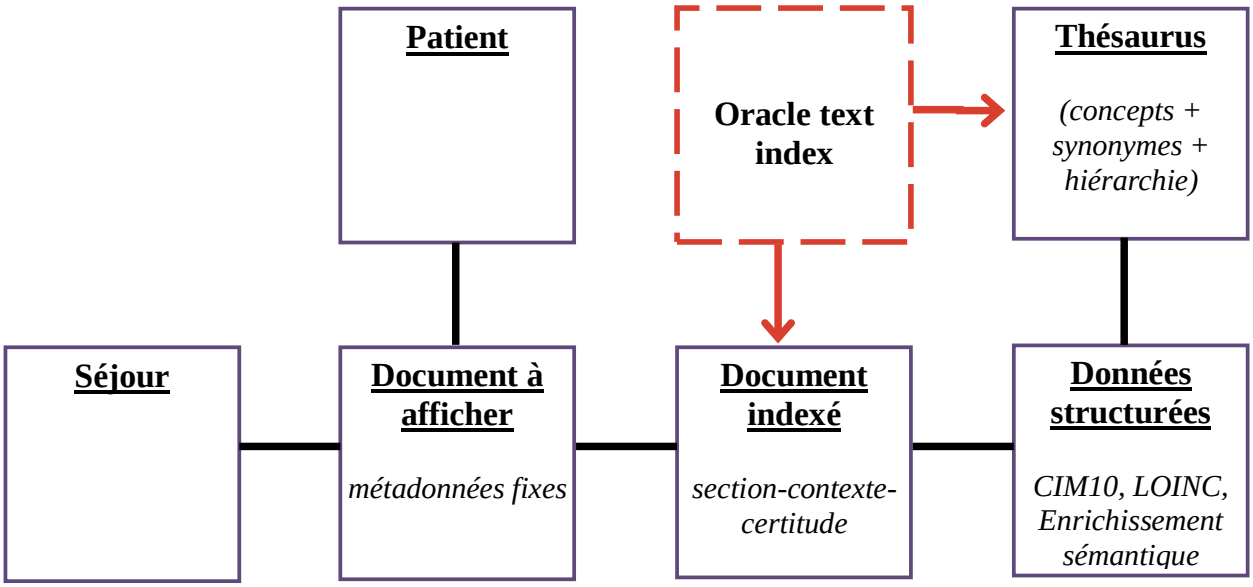


*Figure 1 : schéma relation de l'entrepôt de données*

## 3.6 Intégration au moteur de recherche

Le moteur de recherche multicritère (***Figure 2***) permet de créer plusieurs sous requêtes atomiques. Chaque sous requête peut être une recherche plein texte, une recherche sur donnée structurée (CIM10, LOINC) ou une recherche sur donnée démographique (sexe du patient, âge). L'intersection entre ces sous requêtes se fait soit sur un séjour du patient, soit sur l'ensemble du dossier patient. L'utilisateur peut préciser si une sous requête correspond à un critère d'exclusion ou d'inclusion, et éventuellement sur un contexte particulier (s'il a été détecté : motif, diagnostic etc.). La recherche plein texte exploite le module Oracle Text fourni par défaut dans Oracle 11g. Il permet d'utiliser des opérateurs booléens (or, and, not) ainsi que des opérateurs spécifiques (near, fuzzy etc.). Dorénavant l'utilisateur peut préciser si la recherche peut se faire sur les antécédents familiaux, sur le patient ou sur l'intégralité du compte rendu. Par défaut, le moteur prend en compte la détection de la négation en soustrayant du résultat les patients dont le terme recherché est présent dans le texte dont la certitude est à 0, c'est à dire considéré comme une négation. L'utilisateur peut désactiver ce critère s'il veut obtenir tous les documents qui parlent du concept même avec une négation.

L'enrichissement sémantique est aussi optionnel pour l'utilisateur, il permet d'étendre la recherche non seulement aux concepts synonymes mais aussi aux concepts fils : par exemple la requête « insuffisance cardiaque » permettra de retrouver les valvulopathies. Vu que l'enrichissement sémantique est relié au triplet texte-contexte-certitude, si le texte contient « le patient n'a pas de

valvulopathie», ce ne sera pas un critère d'exclusion du résultat, parce que cela ne sous entend pas qu'il n'a pas d'autres formes d'insuffisance cardiaque dans une autre partie du texte. La relation de subsomption n'est alors pas utilisée dans ce cas.

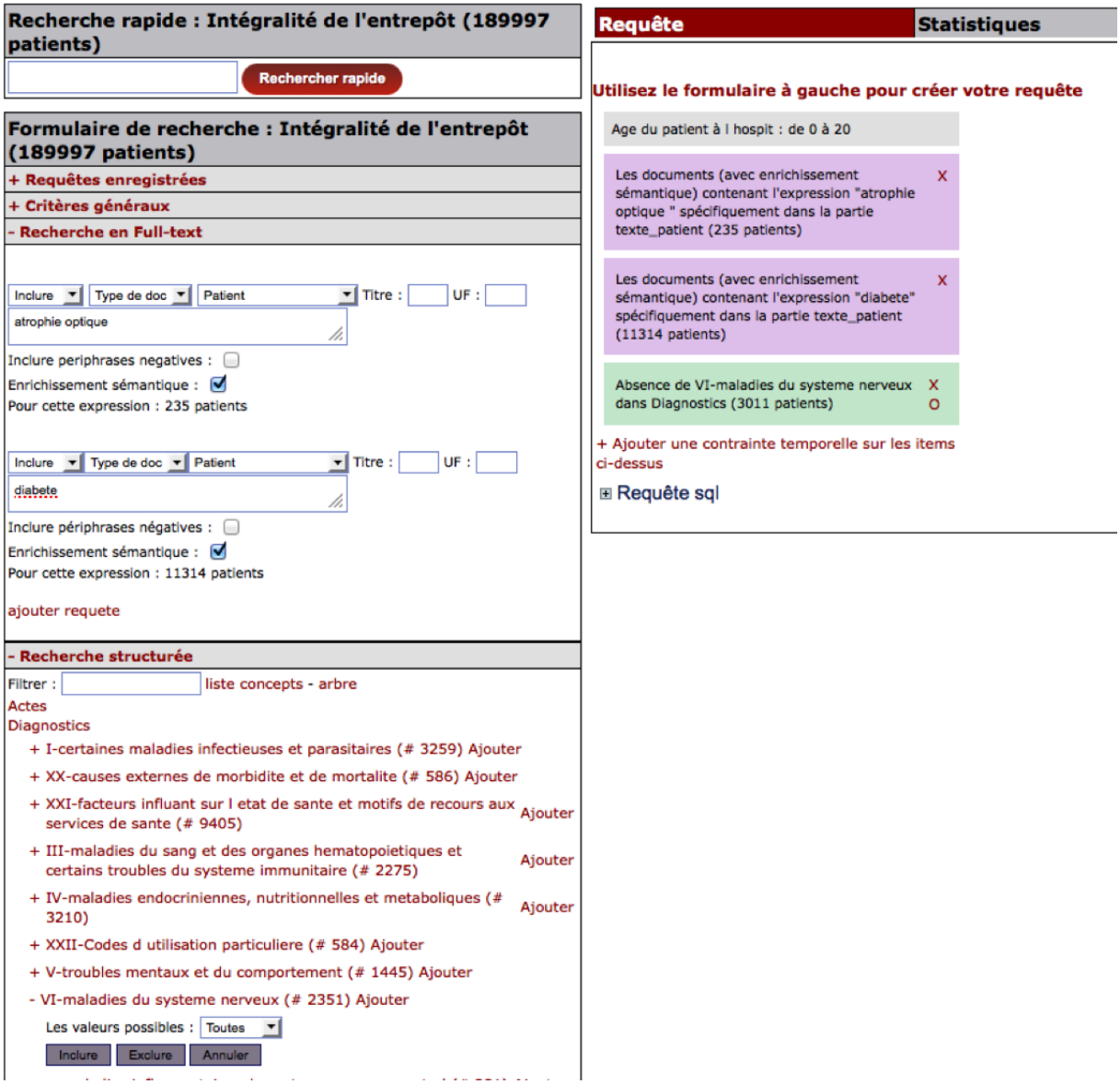


*Figure 2 : Moteur de recherche*

## 3.7 Evaluation

Nous avons tout d'abord réalisé une évaluation intrinsèque de la détection de la négation et des antécédents familiaux associée à l'enrichissement sémantique. Pour cela nous avons exécuté notre algorithme sur 130 comptes rendus médicaux. Nous obtenons un tableur qui donne pour chaque phrase les concepts extraits, la notion de négation ou d'affirmation et le contexte (patient ou famille). Les 1700 lignes du tableur ont été évaluées manuellement afin de calculer la précision et le rappel pour la détection de la négation d'une part, et la détection du contexte d'autre part.

Pour évaluer l'apport dans le moteur de recherche, nous avons exécuté notre algorithme sur l'intégralité de l'entrepôt de données de l'hôpital Necker Enfants Malades : 189 997 Patients et 653 047 documents. Les documents sont issus de plusieurs sources différentes (CR hospitalisation, consultation, radiologie, foetopathologie, les résumés PMSI sont volontairement exclus) et plusieurs formats (compte rendu Word, compte rendu HTML, compte rendu texte). Les textes sont dé-identifiés. Nous avons testé le moteur de recherche sur 3 requêtes plein textes : « crohn and diabete », « NPHP1 » et « lupus and insuffisance renale». Une technicienne médicale a évalué le

nombre de patients vrais positifs, faux positifs pour les 3 requêtes avec et sans la détection de la négation, puis avec la détection des antécédents familiaux. Pour évaluer les résultats, la technicienne médicale utilise l'interface web de l'entrepôt de données permettant de vérifier rapidement le contexte des termes retrouvés au survol de la souris (***Figure 3***). En cliquant sur le document elle peut aussi accéder à l'intégralité du compte rendu afin de vérifier en cas de doute s'il s'agit de vrai positif ou non. Pour le calcul du rappel, nous considérons que le moteur de recherche plein texte sans prise en compte de la négation ni des antécédents familiaux renvoie l'intégralité des vrais positifs. La technicienne médicale génère ainsi le gold standard pour une requête en séparant les vrais positifs des faux positifs. A partir de ces données nous déterminons la précision et le rappel pour la détection de la négation puis pour la détection des antécédents familiaux.

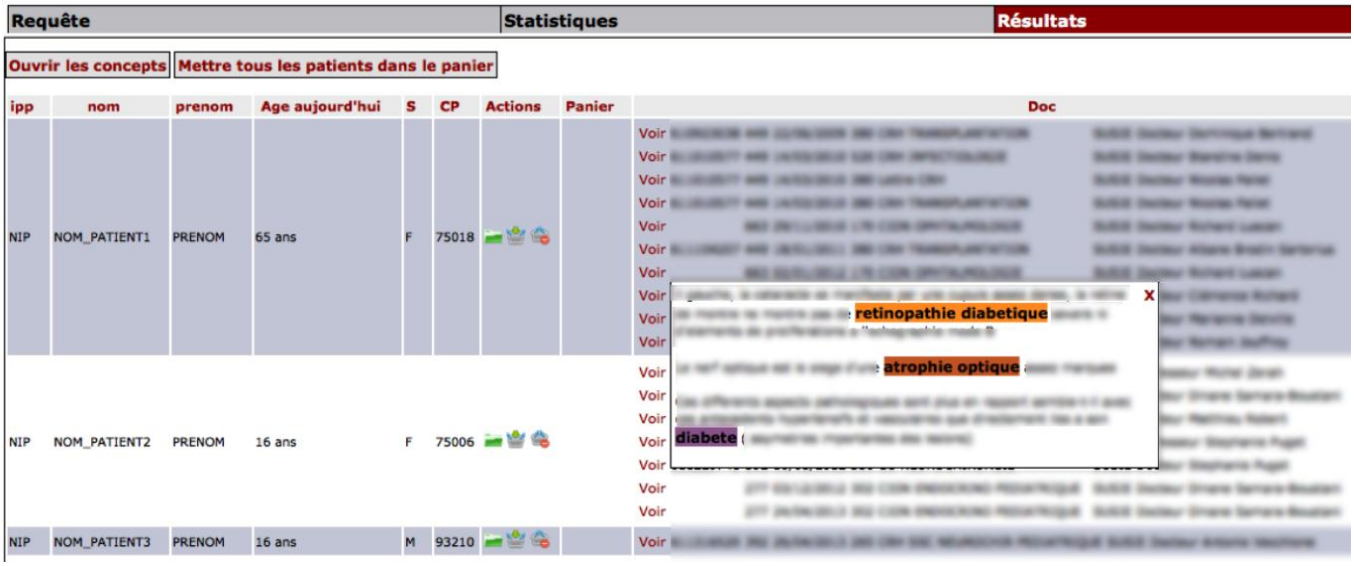


*Figure 3 : Aperçu du résultat*

# 4 Résultats

L'évaluation intrinsèque de la détection de la négation et du contexte associée à l'enrichissement sémantique est représentée dans les tableaux 1 et 2. Le tableau 1 présente la détection du contexte d'un concept.

Le calcul de la précision = Vrais Positifs / (Vrais Positifs + Faux Positifs).

Le calcul du rappel ou sensibilité = Vrais Positifs / (Vrais Positifs + Faux Négatifs).

La spécificité = Vrais Négatifs / (Vrais Négatifs + Faux Positifs).

En considérant l'Histoire familiale comme l'élément positif, nous obtenons une précision de 95,2%, un rappel de 94,4% et une spécificité de 98,7%.

*Tableau 1 : Evaluation de la détection du contexte*

| Contexte | Vrai | Faux |
|---|---|---|
| Histoire familiale | 338 (VP) | 17 (FP) |
| Patient | 1325 (VN) | 20 (FN) |
| Total | 1663 | 37 |

Le tableau 2 représente la détection de la négation et de l'affirmation associée au concept. En

considérant la négation comme l'élément positif, nous obtenons une précision de 95,6%, un rappel de 92,8% et une spécificité de 99,1%

*Tableau 2 : Evaluation de la détection de la négation*

| Contexte | Vrai | Faux |
|---|---|---|
| Négation | 283 (VP) | 13 (FP) |
| Affirmation | 1382 (VN) | 22 (FN) |
| Total | 1665 | 35 |

Sur les 653 047 documents contenus dans l'entrepôt de données, nous avons pu extraire des antécédents familiaux sur 110 774 documents. Le tableau 3 montre la répartition du nombre de textes sur le couple contexte-certitude. Le tableau 4 montre le nombre de concepts UMLS extraits des textes par contexte et certitude. Le niveau de certitude à 1 correspond à l'intégralité du texte, le niveau de certitude à 0 correspond uniquement aux parties de phrases exprimant la négation.

*Tableau 3 : Nombre de documents par contexte et certitude*

| Contexte | Certitude | Nombre de textes |
|---|---|---|
| Texte Intégral | 1 | 653 047 |
| Texte Intégral | 0 | 540 147 |
| Patient | 1 | 653 047 |
| Patient | 0 | 541 217 |
| Antécédents familiaux | 1 | 110 774 |
| Antécédents familiaux | 0 | 5 780 |

*Tableau 4 : Nombre de concepts UMLS par contexte et certitude*

| Contexte | Certitude | Nombre de concepts |
|---|---|---|
| Texte Intégral | 1 | 9 027 310 |
| Texte Intégral | 0 | 1 663 988 |
| Patient | 1 | 8 947 279 |
| Patient | 0 | 1 354 236 |
| Antécédents familiaux | 1 | 80 031 |
| Antécédents familiaux | 0 | 4 032 |

Les tableaux 5, 6, 7 indiquent le nombre de patients Vrai Positifs (VP), Faux Positifs (FP), Vrai Négatifs (VN), Faux Négatifs (FN), la précision et le rappel pour chacune des requêtes dans 3 cas de figure : la recherche plein texte sans prise en compte de la négation ou des antécédents familiaux (texte intégral), la recherche plein texte avec exclusion de la négation (texte intégral - négation), la recherche plein texte avec exclusion de la négation et exclusion des antécédents familiaux (texte intégral - négation - antécédents familiaux).

L'ajout du module gérant la négation apporte un gain de précision (de 0,75 à 0,83 pour « lupus et insuffisance renale »), même si pour la requête « crohn and diabete » il y a une perte du rappel (de 1 à 0,83). Ceci s'explique par une tournure négative n'exprimant pas de négation et que nous n'avions pas prévu dans les exceptions : « je ne reviens pas sur le crohn ». La perte du rappel lors de la prise en compte des antécédents familiaux pour la recherche sur le lupus et l'insuffisance rénale (de 1 à 0,95) s'explique par l'utilisation au sein d'une même phrase du patient et de son frère, l'un ayant la maladie l'autre non. La phrase a alors été considérée comme relatif aux antécédents familiaux et non au patient.

La détection des antécédents familiaux augmente aussi la précision sans modifier le rappel dans les 3 tests effectués, excepté pour la recherche du gène « NPHP1 » qui n'a pas été testé dans les familles des patients. On voit par ailleurs que la précision de la requête « NPHP1 » reste faible malgré une amélioration de la précision de 0,42 à 0,52. Dans notre expérience à l'Hôpital Necker, les utilisateurs qui requêtent avec un nom de gène recherchent les patients avec une mutation sur ce gène. Nous avons donc considéré dans notre évaluation que le doute et la prescription du test génétique sont des faux positifs. Si nous considérons ces éléments comme vrai positif nous améliorons la précision à 0,81 pour « NPHP1 » comme indiqué dans le tableau 8.

*Tableau 5 : Précision et rappel pour « lupus and insuffisance renale »*

| Lupus and insuffisance renale | VP | FP | VN | FN | Précision | Rappel |
|---|---|---|---|---|---|---|
| Texte Intégral | 77 | 26 | 0 | 0 | 0,75 | 1,00 |
| Texte Intégral - négation | 77 | 16 | 10 | 0 | 0,83 | 1,00 |
| Texte Intégral - négation - antécédents familiaux | 73 | 7 | 19 | 4 | 0,91 | 0,95 |

*Tableau 6 : Précision et rappel pour « Crohn and diabete »*

| Crohn and diabete | VP | FP | VN | FN | Précision | Rappel |
|---|---|---|---|---|---|---|
| Texte Intégral | 12 | 82 | 0 | 0 | 0,13 | 1,00 |
| Texte Intégral - négation | 10 | 57 | 25 | 2 | 0,15 | 0,83 |
| Texte Intégral - négation - antécédents familiaux | 10 | 10 | 72 | 2 | 0,50 | 0,83 |

*Tableau 7 : Précision et rappel pour « NPHP1 »*

| NPHP1 | VP | FP | VN | FN | Précision | Rappel |
|---|---|---|---|---|---|---|
| Texte Intégral | 11 | 15 | 0 | 0 | 0,42 | 1,00 |
| Texte Intégral - négation | 11 | 10 | 5 | 0 | 0,52 | 1,00 |
| Texte Intégral - négation - antécédents familiaux | 11 | 10 | 5 | 0 | 0,52 | 1,00 |

*Tableau 8 : Précision et rappel pour « NPHP1 » avec doute et prescription*

| NPHP1 (doute et prescription) | VP | FP | VN | FN | Précision | Rappel |
|---|---|---|---|---|---|---|
| Texte Intégral | 17 | 9 | 0 | 0 | 0,65 | 1,00 |
| Texte Intégral - négation | 17 | 4 | 5 | 0 | 0,81 | 1,00 |
| Texte Intégral - négation - antécédents familiaux | 17 | 4 | 5 | 0 | 0,81 | 1,00 |

# 5 Discussion

L'évaluation intrinsèque de la détection du contexte familiale montre une précision (95,2%), un rappel (94,4%) et une spécificité (98,7%) relativement élevés. Nos scores sont sensiblement inférieurs aux résultats de Goryachev & al. [9] avec une sensibilité de 97,2% et une spécificité de 99,7%. Notre rappel est un peu supérieur au rappel de Friedlin et al. [8] de 93%, mais notre précision est légèrement inférieure (97%). Notre évaluation diffère de celle de Goryachev et al parce qu'ils effectuent l'analyse sur une extraction automatique de la section du compte rendu qui concerne l'histoire familiale. Friedlin évalue lui aussi son algorithme uniquement sur les sections concernant l'histoire familiale, mais avec une détection exacte des sections car précisément labélisées comme histoire familiale. Notre évaluation porte sur l'intégralité de chaque compte rendu. De plus, leur algorithme repose sur la langue anglaise alors que la nôtre est spécifique au français.

L'évaluation intrinsèque de la détection de la négation montre une précision (95,6%), un rappel (92,8%) et une spécificité (99,1%) relativement élevés. Les résultats obtenus par Deléger et Grouin [7] montrent une précision et un rappel respectivement de 84,4% et 83,3% pour NegEx, et 88,8% et 84,6% pour leur adaptation au français de NegEx.

L'apport de notre méthode pour améliorer la qualité du moteur de recherche est variable suivant les requêtes. On voit toutefois qu'il y a à chaque fois un gain de la précision avec peu ou pas de perte de rappel. Les erreurs sont généralement liées à un problème de découpage des phrases lors du nettoyage des textes. Si une phrase se termine sans ponctuation par un saut de ligne et qu'à la ligne suivant la 1ere lettre du mot est en majuscule, l'algorithme considère que ce sont 2 phrases distinctes. Mais cette méthode a ses limites pour les noms propres qui commencent par une majuscule. En effet, pour Crohn, certaines phrases sont découpées de telle manière que le terme Crohn est précédé d'un saut de ligne :

« La sœur du patient a une maladie de↵
Crohn »

L'algorithme considère alors ces deux phrases distinctes : « La sœur du patient a une maladie de » et « Crohn ». L'antécédent familial ne prend alors pas en compte « Crohn » qui reste associé au patient.

L'évaluation nous montre aussi que certains résultats Faux Positifs sont dus à une expression du doute ou d'un examen à réaliser. Par exemple, les médecins font mention du gène NPHP1 non par pour exprimer le fait qu'il est muté ou non muté mais pour émettre une suspicion de mutation ou une prescription d'examen génétique sur ce gène.

Nous pensons pouvoir encore améliorer la qualité de la détection de la négation en ajoutant de nouvelles règles et exceptions découvertes au cours de l'évaluation. Nous souhaitons ajouter un niveau intermédiaire de certitude permettant d'exprimer la suspicion ou le doute avec un score à 0,5. Nous prévoyons de développer un algorithme spécifique aux informations génétiques afin de discriminer les gènes effectivement mutés des tests génétiques prescrits et des suspicions de mutations (doute).

# 6 Conclusion

Nous avons pu intégrer dans l'entrepôt de données de l'hôpital Necker Enfants Malades un algorithme permettant de diminuer le bruit de la recherche plein texte dans les comptes rendus médicaux en extrayant automatiquement les antécédents familiaux des patients et les sous phrases exprimant une négation. L'interface utilisateur permet de choisir le niveau de spécificité souhaité

en fonction de la rareté ou de la complexité des requêtes effectuées. L'extraction des concepts associée à la détection de la négation et des antécédents familiaux nous permettra de travailler sur l'enrichissement des descriptions phénotypiques [11] des patients porteurs d'une anomalie génétique.

## Remerciements

Nous remercions chaleureusement Malika-Anaïs Boujemaoui pour avoir évalué l'amélioration de la précision du moteur de recherche par la détection de la négation et des antécédents familiaux.

## Références

[1] Cuggia M, Bayat S, Garcelon N, Sanders L, Rouget F, Coursin A, Pladys P. A full-text information retrieval system for an epidemiological registry. Stud Health Technol Inform. 2010;160(Pt 1):491-5.

[2] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, Happe A, Duvauferrier R. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584-8.

[3] Humphreys, B.L., et al., The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc, 1998. 5(1): p. 1-11.

[4] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries", 2001; J Biomed Inform, vol. 34, no. 5, pp. 301{10,.

[5] S. Goryachev, M. Sordo, Q.T. Zeng, and L. Ngo. Implementation and evaluation of four different methods of negation detection, 2006; Technical report, DSG

[6] Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc. 1999 Sep-Oct; 6(5):393-411.

[7] L. Deléger & C. Grouin. Detecting Negation of Medical Problems in French Clinical Notes. IHI'12 Proceedings of 2nd ACM SIGHIT International Health Informatics Symposium. 2012.

[8] Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.

[9] Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc. 2008 Nov 6:247-51.

[10] N. Lewis, D. Gruhl, and H. Yang, "Extracting family history diagnosis from clinical texts," in International Conference on Bioinformatics and Computational Biology, March 2011.

[11] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010 May 1;26(9):1205-10. doi: 10.1093/bioinformatics/btq126. Epub 2010 Mar 24.

### Adresse de correspondance

Institut Imagine, 24 boulevard Montparnasse, 75015 Paris

nicolas.garcelon@institutimagine.org