F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED Internationale

Suzanne Pereira[1, 2, 3], Philippe Massari[1], Antoine Buemi[4], Badisse Dahamna[1], Elisabeth Serrot[3], Michel Joubert[2], Stéfan J. Darmoni[1]

[1]CISMeF, CHU de Rouen, & TIBS, LITIS EA 4108, Université de Rouen, France
[2]LERTIM, Université de la Méditerranée, Marseille, France
[3]Vidal, Issy les Moulineaux, France
[4]AP-HP, Paris, France

Abstract

Background: SNOMED is becoming a major health terminology to index electronic medical records. Most of developed countries have chosen SNOMED CT. France has chosen SNOMED International which is already translated in French. We developed the F–MTI tool, a generic automatic indexing tool able to index documentation in several health terminologies written in French (CCAM, TUV) or translated in French (MeSH, ICD-10, SNOMED International). Objective: The objective of the paper is to evaluate the discharge summaries indexing in SNOMED performed by F-MTI and compare it to the one performed by Snocode (a commercial Canadian product). Methods: We compared F–MTI and Snocode on a corpus of 100 discharge summaries from the Rouen University Hospital. Results: The results showed a Hooper's measure of 32.9% comparing the two sets of SNOMED codes. With the help of a SNOMED-ICD10 mapping we could compare in terms of diagnosis these two sets with a manual coding. We obtained a precision of 6.1% for Snocode and 4.4% for F-MTI and respectively a recall of 24.7% and 27.0%. Conclusion: Snocode and F-MTI indexing are as close as two manual indexing can be, in terms of precision and recall. They also provided close results in terms of diagnosis.

Keywords

Abstracting and Indexing/methods; Algorithms; Information Storage and Retrieval/methods; Evaluation Study, France; Natural Language Processing; Systematised Nomenclature of Medicine; medical records; international classification of diseases

1 Introduction

Représenter les données du dossier médical informatisé à l'aide de terminologies standardisées permet de faciliter les pratiques médicales, de rechercher plus facilement de l'information et de mieux gérer les coûts [1, 2].

La recherche en informatique médicale a montré que la SNOMED (nomenclature systématique de médecine humaine et vétérinaire) est la terminologie la plus adaptée à l'indexation des informations du dossier médical [1, 2]. En 2006, la France s'approprie les droits de la version en français de la SNOMED Internationale (SNOMED 3.5 VF) avec

l'objectif d'en faire l'outil d'indexation des dossiers médicaux.

Ce choix peut paraître anachronique alors que plusieurs pays choisissent une version plus récente, la SNOMED CT (Clinical Terms) [3], mais se justifie par la non disponibilité d'une traduction française de cette version[1]. La SNOMED 3.5 VF bénéficie d'une validation partielle de la traduction grâce aux travaux du consortium VUMeF[2] (2003-2006) financé par le programme RNTS. La SNOMED 3.5 et la SNOMED CT sont très proches sémantiquement (la première est incluse à 91% dans la seconde), et une compatibilité à rebours existe entre les deux versions : la convergence avec la SNOMED CT, lorsque celle-ci sera disponible en français, est par conséquent facilement envisageable.

La SNOMED [4] est, à l'origine, une extension à l'ensemble de la médecine du concept développé par le College of American Pathologists avec la SNOP (Nomenclature Systématique d'anatomie Pathologique, 1965) [5], À ce jour, la version 3.5 contient plus 150 000 termes répartis en 11 axes de description. Un axe regroupe les termes d'un sous domaine (par exemple, l'axe D correspond aux diagnostics). Tous les axes peuvent être combinés pour former un concept SNOMED. La SNOMED 3.5 est traduite en 11 langues, la version anglaise étant incluse dans l'UMLS (Unified Medical Language System [6]).

Quel que soit le système proposé, les médecins sont réticents à pratiquer une indexation précise et exhaustive des dossiers médicaux, essentiellement du fait du grand nombre de codes à rechercher et de la complexité des référentiels proposés. Des systèmes d'aide à l'indexation et au codage sont nécessaires.

Certaines solutions existent pour la SNOMED CT utilisant les variations lexicales anglaises de l'UMLS [7, 8]. Pour la SNOMED RT (Reference Terminology [9]) et 3.5, elles utilisent les transcodages vers la SNOMED 3.5 [10, 11]. Certains produits commerciaux sont aussi disponibles : Language Engine-LE[3], CoPath Plus[4], 3M. DialeCT. Encoder[5] et Snocode[6].

À notre connaissance, le seul outil existant pour la SNOMED 3.5 VF est Snocode. C'est un outil d'indexation automatique interactif ou autonome destiné à l'indexation des textes cliniques proposant des codes SNOMED 3.5, ainsi que des termes provenant d'autres terminologies comme la Classification Internationale des Maladies 9^{ième} et 10^{ième} révisions (CIM9 et CIM10 [12]), la CCAM (Classification Commune des Actes Médicaux)[13] et la SNOMED CT. L'outil Snocode est disponible en plusieurs langues, incluant l'anglais et le français. En France, depuis 2000, Snocode est intégré au système de gestion des services d'anatomie pathologique des hôpitaux de Colmar et de Mulhouse.

Nos travaux sur l'indexation du dossier médical français débutent en 2005. Une étude préliminaire compare les codes CIM10 produits par un système combinant un extracteur MeSH (Medical Subject Heading [14]) et un transcodage entre le MeSH et la CIM10 à celle produite par Snocode combiné avec le même transcodage sur le même corpus [15]. Les résultats montrent que l'utilisation de la SNOMED avec Snocode à la place du MeSH aboutit à une précision de 25% comparé à un codage descriptif manuel (vs. 4% pour notre système) et un rappel de 46% comparé à un codage descriptif (vs. 68% pour notre système).

---

[1] Une traduction partielle en français de la SNOMED CT est en cours au Canada (URL : http://www.infoway-inforoute.ca/).
[2] Vocabulaire unifié médical français
[3] http://www.healthlanguage.com
[4] http://www.misyshealthcare.com
[5] http://www.3m.com
[6] http://www.medsight-info.com

Le développement d'un système appelé F-MTI (French Multi-Terminology Indexer[7]) suit ces premiers travaux. F-MTI est un outil d'indexation automatique générique capable d'indexer des documents à l'aide de plusieurs terminologies. Nous avons déjà implémenté cinq terminologies françaises ou traduites en français : SNOMED 3.5, CIM10, MeSH, CCAM et TUV[8]. Il indexe directement et indirectement un document en utilisant les transcodages. Nous avons évalué cet outil lors de l'indexation de comptes rendus à l'aide de la CIM10. F-MTI aboutit à une précision de 3.7% et un rappel de 32.9% comparé à une indexation manuelle CIM10 [17].

Le but de ce travail est de confirmer nos objectifs précédents d'indexation multi-terminologique et de montrer que F-MTI peut aussi indexer un document en SNOMED. Dans ce travail, nous comparons les deux propositions d'indexation SNOMED de F-MTI et de Snocode produites à partir d'un échantillon de 100 comptes rendus d'hospitalisation. L'indexation manuelle de ces documents est considérée comme la référence.

2 Matériel et méthodes

2.1 Snocode

Snocode3 incorpore les technologies propriétaires de stockage d'information et d'indexation en langage naturel de la société Medsight. Il utilise des techniques de transcodage et de synonymie pour comparer des séquences de mots d'un texte clinique aux termes SNOMED restructurés afin de faciliter et de rendre plus rapides les comparaisons. La stratégie de base consiste à ne retenir que les correspondances exactes et les plus longues. En même temps, une analyse syntaxique est utilisée pour les transformations singulier et pluriel. Les résultats peuvent être paramétrés. Dans notre étude, nous avons choisi les options : codes SNOMED et codes CIM10. Snocode est utilisé en mode autonome (sans action humaine). Le mode « une phrase à la fois » est sélectionné. Pour la comparaison au système F-MTI, tous les codes pour chaque phrase sont groupés et les redondances sont éliminées.

2.2 F-MTI

La première version de F-MTI est créée en 2006. Elle indexe un document à l'aide des terminologies implémentées : MeSH, SNOMED, CIM10, CCAM et TUV. Puis toutes les terminologies sont projetées vers les terminologies désirées par l'utilisateur (ici nous sélectionnons la CIM10 et la SNOMED) par le biais des transcodages. La plupart de ces transcodages proviennent du méta-thesaurus de l'UMLS. Certains sont développés par l'équipe CISMeF (transcodage de la CCAM vers le MeSH) ou par la société Vidal (transcodage TUV vers la CIM10). Cette méthode permet de retrouver plus de termes en identifiant plus de formes textuelles pour les termes des différentes terminologies. En effet, chaque terminologie a ses propres termes, synonymes et variations pour un même concept, les concepts équivalents dans différentes terminologies étant reliés par des liens de transcodage. Cette approche multi-terminologique a été définie pour la première fois par la US National Library of Medicine avec MTI (MeSH Terminology Indexer [16]). À notre connaissance, F-MTI est le premier outil multi-terminologique développé pour une autre langue que l'anglais. Cette réalisation est étroitement liée à la présence du français dans 10 terminologies de santé incluses dans l'UMLS (sans parler d'autres terminologies qui ne le sont toujours pas, comme la CCAM et le TUV).

[7] En rappel du projet MTI de la US National Library of Médicine [16]

[8] Terminologie Unifiée du Vidal décrivant les propriétés thérapeutiques des médicaments.

Dans notre étude, nous utilisons l'algorithme du sac de mots implémenté dans F-MTI. Voici son fonctionnement :

- D'abord, le document est converti en texte et les phrases sont extraites avec la rubrique correspondante grâce à un transducteur NOOJ [19].
- Ensuite, chaque phrase est normalisée (les accents sont éliminés, minuscules…) et découpée en mots. Les mots vides et non pertinents (car non inclus dans un terme) sont éliminés. Chaque mot est ensuite désuffixé (stemming en anglais). F-MTI est aussi capable de prendre en compte certains éléments du contexte pour une phrase. Le stème correspondant à la rubrique « antécédents » est ajouté pour les phrases appartenant à rubrique correspondante dans le compte rendu médical.
- Le sac de mots ainsi obtenu est mis en correspondance indépendamment de l'ordre des mots avec les termes des terminologies MeSH, SNOMED, CIM10, CCAM et TUV traités de la même manière. Les règles pour la mise en correspondance sont : tous les termes correspondant à au moins un stème du sac sont retrouvés et les correspondances les plus longues sont privilégiées par rapport aux plus petites.
- Puis, les termes d'indexation candidats sont restreints aux termes SNOMED 3.5 les plus proches sémantiquement en utilisant les transcodages entre concepts. Les transcodages sont : MeSH->SNOMED 3.5 extrait de l'UMLS2007AA et une version du transcodage CIM10 vers SNOMED incluant le transcodage de l'UMLS2007AA et un transcodage français non validé inclus dans la version en français de la SNOMED. Une étude a débuté pour le valider [20].
- Finalement, tous les termes SNOMED et CIM10 extraits du document sont groupés. Les redondances sont éliminées, et les termes plus précis sont gardés.

2.3 Comparaison des deux systèmes : F-MTI et Snocode

Initialement, l'objectif de cette étude est de comparer les deux systèmes à l'aide d'une indexation manuelle SNOMED réalisée par un expert sur un échantillon de comptes-rendus d'hospitalisation. Cette indexation descriptive (hors problèmes de codage PMSI) aurait été la première expérience d'indexation manuelle à l'aide de la SNOMED en France.

Toutefois, cette indexation manuelle en SNOMED s'avère trop consommatrice en temps : un jour de travail est nécessaire à l'expert pour indexer un compte rendu de 3 pages. Ce temps était partagé entre la reformulation du texte original, l'élimination des ambiguïtés, la reconnaissance des acronymes inhabituels, l'analyse du contexte, la reconnaissance des concepts et l'analyse des synonymes SNOMED. Lorsque le texte est correctement structuré et formulé sans ambigüités et non-dits, le temps d'indexation est beaucoup plus court. La nécessité de disposer de documents structurés avec une rigueur descriptive est ainsi réaffirmée !

Après ce constat, la projection des codes SNOMED vers une terminologie moins complexe pouvant être indexée manuellement semble être une bonne façon de comparer ces indexations, en termes de diagnostics tout au moins. La terminologie CIM10, utilisée dans les hôpitaux en France dans un but médico-économique, est retenue pour notre étude. Cette méthode indirecte reproduit la nécessité de coder les comptes rendus en CIM10 (voir figure 1) pour le PMSI en France [21].

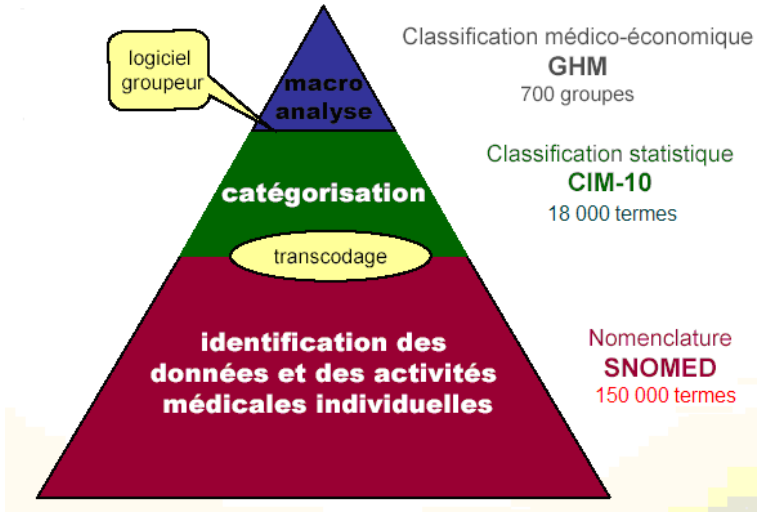

Figure 1 : Les niveaux de l'information médicale

Les deux ensembles de codes SNOMED produits par F-MTI et Snocode sont comparés à l'aide de quelques mesures sans aucune référence. Puis les deux ensembles résultant de la projection des codes SNOMED en codes CIM10, sont comparés en utilisant une indexation manuelle descriptive à l'aide de la CIM10 réalisée par un expert.

Le processus de projection de la SNOMED vers la CIM10 est différent dans les deux systèmes. Snocode utilise un transcodage réalisé par le SFINM (Secrétariat francophone international de nomenclature médicale) et F-MTI utilise une combinaison de celui-ci avec celui inclus dans l'UMLS2007AA. Nous disposons en interne de celui de F-MTI, c'est donc celui-ci qui est utilisé préférentiellement dans la suite de ce travail.

C'est pourquoi nous présentons deux évaluations : une pour les différents transcodages et l'autre en utilisant seulement le transcodage de F-MTI.

2.4 Corpus choisi

Un échantillon de 100 comptes rendus rédigés par des médecins du CHU de Rouen en 2005 est utilisé pour cette évaluation. 50 comptes rendus proviennent du service de cardiologie et 50 de celui de pneumologie. Ces choix découlent du fait qu'il s'agit des domaines de prédilection de notre expert en indexation médicale (PM). Les dossiers sont extraits du système de dossier médical électronique du CHU de Rouen (1 080 384 patients et 182 808 comptes rendus en 2005). Un compte rendu médical décrit l'histoire de la maladie, les actes effectués, les examens et la prescription de médicaments. Après avoir lu chaque compte rendu, l'expert est invité à les indexer de manière descriptive à l'aide de la CIM10 pour former notre indexation de référence.

2.5 Mesures d'évaluation

Nous utilisons la mesure de Hooper afin de comparer les deux ensembles de codes SNOMED [24]. Cela mesure la consistance de l'indexation entre deux indexeurs. Nous avons considéré F-MTI et Snocode comme deux indexeurs différents potentiels [22]. La formule de Hooper est :

$$H=100*C/(F+S+C) \quad (1)$$

C= nombre de termes extraits par Snocode (S) et F-MTI (F); F= nombre de termes extraits par F mais pas par S; S= nombre de termes extraits par S mais pas par F.

Nous calculons aussi la couverture de chaque indexation l'une par rapport à l'autre.

Puis, nous mesurons la précision, le rappel et la F-mesure afin de comparer les deux indexations à la référence. La précision et le rappel sont des mesures usuelles en sciences de l'information. La précision est le ratio du nombre de codes pertinents extraits (codes proposés par le système et par la référence) sur le nombre total de codes extraits (par le système). Le rappel est le ratio du nombre de codes pertinents extraits sur le nombre total de codes de la référence.

La F-mesure est la moyenne pondérée harmonique de la précision et du rappel [23]. La formule de la F-mesure est :

$$F=1/(\alpha*(1(1/P)+\alpha*(1/R)) \quad (2)$$

P=précision ; R=rappel ; Dans notre cas $\alpha=1$

3 Résultats

Les résultats de cette étude sont présentés dans les tableaux 1, 2 et 3.

Tableau 1 : Comparaison des outils Snocode et F-MTI en ne considérant que les codes SNOMED

	Nombre de codes S NOMED par CRH
Snocode	54.9
F-MTI	100.3
Pourcentage des codes extraits par F-MTI recouvrant les codes de Snocode	29.9%
Pourcentage des codes extraits par Snocode recouvrant les codes de F-MTI	51.5%
Mesure de Hooper	31.3%

Le tableau 1 montre que Snocode extrait moitié moins de codes que F-MTI (54.9 vs. 100.3). La moitié des codes extraits par Snocode sont aussi extraits par F-MTI.

Tableau 2 : Comparaison des outils Snocode et F-MTI en ne considérant que les codes CIM10 produits par différentes projections des codes SNOMED vers les codes CIM10

	Nombre de codes CIM10 par CRH		
Indexation manuelle	4.2		
Snocode	6.5		
F-MTI	26.5		
Pourcentage des codes extraits par F-MTI recouvrant les codes de Snocode			4.4%
Pourcentage des codes extraits par Snocode recouvrant les codes de F-MTI			9.9%
Mesure de Hooper			17.4%
	Précision %	**Rappel %**	**F-measure %**
Snocode	15.0	22.2	17.9
F-MTI	4.4	30.7	3.8

Les tableaux 2 et 3 présentent les résultats de la comparaison des deux outils après transcodage vers la CIM10. Le changement du type de transcodage produit des résultats différents. Le nombre moyen de codes extraits par compte-rendu passe de 7 à 17 codes extraits par Snocode (vs F-MTI 26.5 codes et 4.2 pour l'indexeur humain).

F-MTI trouve plus de codes que Snocode et que l'expert, ce qui donne une précision très faible 4.4%. Snocode produit une meilleure précision 15% et 6.1% avec le même transcodage. Les scores se rapprochent beaucoup lorsque l'on utilise le même transcodage. F-MTI produit un meilleur rappel (30.7% vs 22.2%) et une plus faible précision (4.4% vs 6.1%) par rapport à Snocode. La F-mesure de Snocode est supérieure à celle de F-MTI : 17.9 et 4.6. Ces résultats sont beaucoup plus proches (voir tableau 3) entre les deux outils quand on utilise les mêmes outils de transcodage.

Tableau 3 : Comparaison des outils Snocode et F-MTI considérant les codes CIM10 produits par la même projection des codes SNOMED vers les codes CIM10

	Nombre de codes CIM10 par CRH		
Indexation manuelle	4.2		
Snocode	17.1		
F-MTI	26.5		
Pourcentage des codes extraits par F-MTI recouvrant les codes de Snocode			4.4%
Pourcentage des codes extraits par Snocode recouvrant les codes de F-MTI			6.1%
Mesure de Hooper			32.9%
	Précision %	**Rappel %**	**F-measure %**
Snocode	6.1	24.7	12.2
F-MTI	4.4	30.7	3.8

4 Discussion

4.1 Comparaison entre les propositions d'indexation de Snocode et de F-MTI

Il n'est pas surprenant que le nombre de codes générés par les deux systèmes varie grandement (moyenne de 55 codes SNOMED pour Snocode contre 100 pour F-MTI ; moyenne de 17 codes CIM10 pour Snocode vs. 26 pour F-MTI). Ces variations sont dues au fait que Snocode se fonde uniquement sur la nomenclature SNOMED alors que F-MTI se fonde sur quatre autres terminologies pour générer des codes SNOMED.

Dans le tableau 1, la mesure de Hooper montre que les deux outils produisent des indexations aussi différentes que peuvent l'être deux indexations humaines (31.3%). À titre de comparaison à la NLM, les indexeurs humains génèrent une mesure de Hooper de 39% pour l'indexation manuelle des articles à inclure dans MEDLINE avec le thésaurus MeSH [24].

D'après les tableaux 2 et 3, nous pouvons suggérer que les principales différences de résultats entre les deux outils sont liées aux différences de transcodage SNOMED-CIM10 utilisés. L'application du même transcodage que celui utilisé par F-MTI, aboutit à une diminution de 8.9% de la précision et une augmentation du rappel de 2.5% pour Snocode.

La projection des codes SNOMED vers la CIM10 montre que, comparé à une indexation

manuelle, Snocode produit une meilleure précision (+2%) et un plus faible rappel (-6%) que F-MTI en termes d'extraction de maladies. Les résultats peuvent être considérés comme assez proches alors que nous comparons un système mono-terminologique de plus de 20 ans d'expérience (Snocode) et un système multi-terminologique de seulement 6 ans et qui peut encore beaucoup évoluer (F-MTI). Sachant que Snocode est un outil déjà commercialisé et en place dans certains hôpitaux, nous pouvons considérer que les résultats obtenus par F-MTI sont relativement satisfaisants.

4.2 Analyse des erreurs

L'analyse de l'indexation produite par F-MTI met en évidence quelques difficultés ou erreurs :

- L'extraction de termes non pertinents pour l'indexation, par exemple les termes de l'axe G de la SNOMED contenant les qualificatifs et termes de relations qui n'ont aucun sens lorsqu'ils ne sont pas reliés aux autres termes SNOMED.
- F-MTI, tout comme Snocode, ne permet pas de post-coordonner des termes appartenant à différents axes de la SNOMED lors de leur indexation. Il n'existe pas de règles d'indexation à ce sujet. Il est donc nécessaire d'implémenter des règles afin d'indexer correctement les comptes rendus médicaux. Cette étape devra être réalisée en post-coordination.
- Certains termes sont incorrectement retrouvés car l'extraction par la méthode du sac de mots ne permet pas de respecter l'ordre des mots. Des améliorations doivent être apportées dont l'implémentation de l'analyse sémantique des phrases.
- Les transcodages relient généralement des concepts de sens strictement équivalents mais parfois des degrés de précision différents peuvent être rencontrés. Les transcodages doivent donc être validés afin d'éliminer les transcodages inadéquats et ainsi faire diminuer le bruit généré par F-MTI.
- Les redondances entre termes extraits : les diagnostics et leurs symptômes ou différentes formes du même diagnostic ou bien encore la manifestation et la maladie initiale. Les relations «symptôme de» et «diagnostic de» sont présents dans la SNOMED CT qui est reliée par des relations de synonymie à la SNOMED 3.5 dans l'UMLS (car reliés aux mêmes concepts UMLS - voir section 2.3.2.3). Un travail réalisé par un doctorant de l'équipe CISMeF [25] permet de transposer les relations «symptôme de» et «diagnostic de» de la SNOMED CT à la SNOMED 3.5. Une future version de F-MTI doit intégrer ces règles et ces relations.
- L’analyse du contexte : antécédents, autre membre de la famille touchée, négations etc... Des améliorations au niveau de l'analyse du contexte, avec par exemple des transducteurs, sont en cours d’implémentation.
- F-MTI ne peut raisonner comme un médecin et, par exemple, associer des informations provenant de différentes parties du texte. Un système de règles peut être utile ici.
- La formulation des textes soumis à l’indexation : il existe un manque de précision des comptes rendus avec, par exemple, des diagnostics non décrits. Les médecins doivent être invités à une plus grande rigueur descriptive de leurs comptes rendus.

4.3 Évaluation

Cette approche d'évaluation consistant à employer un transcodage vers d'autres terminologies moins complexes produit un biais important. Mais, nous partons de l’hypothèse que si un code CIM10 est produit par Snocode et par F-MTI il doit être issu

d’un (ou plusieurs) codes SNOMED 3.5 équivalents, proches ou sémantiquement reliés. Cette évaluation permet donc de comparer en termes de diagnostics, la proximité (et non l’équivalence) des indexations produites par les deux outils.

Une évaluation manuelle par un expert des codes SNOMED 3.5 produits par les deux outils (sans biais provoqué par un transcodage quelconque) est en cours.

Cette approche d'évaluation peut facilement s’appliquer à d'autres évaluations où l'indexation manuelle est difficile, par exemple la SNOMED CT qui est beaucoup plus complexe que la SNOMED 3.5 et qui possède des liens d'équivalences avec la CIM10 dans l'UMLS.

Un expert n'indexe manuellement pas plus de 5 codes par compte-rendu. En revanche, un outil automatique indexe dix fois plus de codes. Ce qui conduit à s’interroger sur la pertinence d’un codage exhaustif et sur la sélection des informations importantes selon l’utilisation qui en est faite. Dans sa pratique clinique, le médecin ne souhaite consulter que les éléments importants comme les maladies en cours et les derniers traitements. Dans le cadre médico-économique, les termes d'indexation sont souvent limités aux codes classants en vue de la tarification. En revanche, dans un contexte de recherche d’information, d'analyse de données ou d'alertes, une extraction complète des concepts présents dans le compte-rendu et décrits par la terminologie utilisée est préférable.

Dans cet esprit, il est prévu de réaliser une évaluation secondaire qualitative des codes extraits par F-MTI en assignant à chaque code, à dire d’expert, une étiquette associée à sa pertinence en matière de recherche d’information dans le dossier patient («pertinent», «non pertinent» et «peu pertinent»).

4.4 Indexation SNOMED : une tâche complexe

La nomenclature SNOMED 3.5 VF contient sept fois plus de termes que la CIM10 et son utilisation est plus complexe du fait de la multiaxialité. Vu le peu de temps qui peut être consacré à l'indexation, il est certain que cette dernière ne peut se concevoir sans une assistance informatique ou une restriction très sévère des termes utilisés. Ces observations se transposent à l'indexation en SNOMED CT, celle-ci renfermant environ 370000 concepts et 1000000 de termes (presque trois fois plus que la SNOMED 3.5) et plus de 1300000 relations (dans sa version 2007).

4.5 Perspectives

Plusieurs améliorations sont prévues pour F-MTI : management des négations, implémentation de dictionnaires et transducteurs. Avec l’intégration de F-MTI dans BIBLIS, un outil d’annotation [26], F-MTI sera capable de souligner manuellement les parties du texte que l’on désire indexer automatiquement (comme dans Snocode). BIBLIS sera aussi capable d’enregistrer les liens entre les parties du texte original et les termes.

F-MTI est un outil générique capable d’indexer n’importe quelle terminologie en français à partir du moment où sa structure est compatible à celle de sa base de données. Celle-ci est créée sur un modèle générique inspiré du méta-thesaurus de l’UMLS. La SNOMED CT pourra facilement être intégrée lorsqu’elle sera traduite en français.

En pratique, nous envisageons d’exploiter les propositions d’indexation automatique produites par F-MTI dans un outil d’aide à l’indexation de documents. Le médecin, indexant son compte-rendu d’hospitalisation, pourra être aidé par les propositions déjà effectuées par l’outil. Ce système pourrait être aussi intégré sans interaction avec le médecin mais ceci demande encore que l’outil soit amélioré.

L'outil indexe un compte rendu en 1,8s en moyenne. Nous envisageons d'expérimenter l'outil en pratique semi-automatique, avec analyse du temps nécessaire à l'indexation manuelle complémentaire et du gain qualitatif.

Le développement de l'extracteur multi-terminologie F-MTI fait partie d'un projet plus important dans l'équipe TIBS du laboratoire LITIS portant sur l'univers multi-terminologique. Trois thèses ont commencé en 2007 et 2008 : deux s'intéressent à la recherche d'information multi-terminologique en contexte [18][9] pour des portails de santé de l'Internet, comme le catalogue CISMeF [14] et pour des dossiers électroniques de patient. La troisième thèse s'intéresse à l'interopérabilité inter et intra terminologies en santé[10].

5 Conclusion

Nous avons développé un indexeur multi-terminologique automatique effectif pour l'indexation en SNOMED 3.5, CIM10, CCAM, TUV et MeSH. La comparaison entre F-MTI et Snocode montre que les outils Snocode et F-MTI produisent des résultats proches. Ceci est encourageant pour notre projet. Après certaines améliorations nous espérons voir F-MTI intégrer un système de dossier patient électronique.

Remerciements

Ces travaux ont été financés par la société Vidal (http://www.vidal.fr) et le projet PSIP FP7-ICT-1-5.2-Risk Assessment en Patient Safety.

Références

[1] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute C, Warren J. Phase II Evaluation of Clinical Coding Schemes : Completeness, Taxonomy, Mapping, Definitions and clarity. *JAMIA* 1997:238-251.

[2] Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc* 2003:699-703.

[3] College of American Pathologists. SNOMED Clinical Terms Guide. January 2006.

[4] Côté RA, Rothwell DJ, Patolay JL, Beckett RS, and Brochu L, eds. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. *College of American Pathologists*, Northfield, 1993.

[5] Côté RA. From SNOP to SNOMED - A Challenge for the Medical Record Librarian. *Bulletin of the Canadian Association of Medical Record Librarians*; vol.5, no1, December 1972.

[6] McCray AT and Nelson SJ. The semantics of the UMLS knowledge sources. *Methods Inf Med* 1995:34(1/2).

[7] Brigl B, Mieth M, Haux R, Glück E. The LBI-method for automated indexing of diagnoses by using SNOMED. Part 2. Evaluation. *Int J Biomed Comput* 1995 Feb:38 (2):101-8.

---

[9] Cette thèse est en partie financée par le projet européen PSIP (Patient Safety Through Intelligent Procedures in Medication), 7ème PCRD

[10] Ce projet est en partie financé par le projet InterSTIS (programme ANR Tecsan 2007)

[8] Long W. Extracting diagnoses from discharge summaries. *AMIA Annu Symp Proc.* 2005:470-4.

[9] Spackman, K.A. and Campbell, K.E. and Côté, R.A. SNOMED RT: a reference terminology for health care. AMIA Annu Symp Proc, 1997, 640-4.

[10] Moore GW, Berman JJ. Automatic SNOMED coding. *Proc Annu Symp Comput Appl Med Care* 1994:225-9.

[11] Sager N, Lyman M, Nhàn NT, Tick LJ.Automatic indexing into SNOMED III: a preliminary investigation. *Proc Annu Symp Comput Appl Med Care* 1994:230-4.

[12] International Statistical Classification of Diseases and Related Health problems: World health Organisation (WHO), 1992.

[13] Hanser S., Zaiss A., Schulz S. Comparison of ICHI and CCAM basic coding system. *Stud Health Technol Inform* 2006:124:795-800.

[14] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J* 2004 Dec:21(4):253-61.

[15] Pereira S., Névéol A., Massari P., Joubert M. and Darmoni S.J. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. MIE2006:845-850

[16] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *MedInfo* 2004.

[17] Pereira S, Massari P, Joubert M, Serrot E, Darmoni SJ. Exploring Multi-terminology Indexing of Discharge Summaries. *MIE2008* poster.

[18] Sakji S. Recherche multi-terminologique de l'information de santé sur l'Internet. CORIA (cinquième édition de la COnférence en Recherche d'Information et Applications), 2008.

[19] Silberztein M. NOOJ's Dictionnary. LTC2005.

[20] Buemi A., Boisvert M. A., Côté R.A. Transcodage SNOMED - CIM-10 : proposition pour une meilleure indexation des dossiers médicaux. *JFIM2002*.

[21] Geoffroy-Perez B, Imbernon E, Gilg Soit Ilg A, Goldberg M. Comparison of the French DRG based information system (PMSI) with the National Mesothelioma Surveillance Program database. *Rev Epidemiol Sante Publique* 2006:54(6):475-83.

[22] Hooper R. Indexer consistency tests—Origin, measurements, results and utilization. Bethesda: *IBM* 1965.

[23] Rijsbergen CJ. Information Retireval. *Butterworths*, London, 1979.

[24] Funk, M. E., Reid, C. A., & McGoogan, L. S. Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc*. 1983:2 (71):176-83.

[25] Merabti T, Pereira S, Lecroq T, Joubert M, Darmoni SJ. Inheritance of SNOMED CT Relations between concepts to two Health Terminologies (SNOMED International and ICD-10). *KR-Med*2008.

[26] Patriarche R, Gedzelman S, Diallo G, Bernhard D, Bassolet C, Ferriol S, Girard A, Mouries M, Palmer P, Simonet A, Simonet M. A Tool for Textual and Conceptual Annotations of Documents. *eChallenges* 2005.

Adresse de correspondance

Suzanne Pereira, Equipe CISMeF, Bibliothèque Médicale, 1 rue de Germont 76031 Rouen Cedex ; Courriel : Suzanne.pereira@chu-rouen.fr