Développement et évaluation d'une méthode de codage automatique des endoscopies digestives

Iris Ternois[1,2], Jean Baptiste Escudié[1], Catherine Duclos[1,2]

[1] HÔPITAL AVICENNE, APHP, Bobigny, France
iris.ternois@gmail.com, jean-baptiste.escudie@aphp.fr, catherine.duclos@aphp.fr

[2] LIMICS, UNIVERSITÉ PARIS 13, INSERM UMRS 1142 Bobigny, France

Résumé : Les endoscopies digestives sont codées manuellement avec la CCAM par les médecins, ce qui requiert du temps et une bonne connaissance de la terminologie. Cette étude propose un codage automatique de ces actes à partir des comptes rendus. Avec une méthode d'apprentissage supervisé, nous classons les endoscopies diagnostiques avec une précision et un rappel moyens de 0.94, sur un corpus initial de 1539 comptes rendus, séparé en un échantillon d'entraînement et un échantillon de test.

Mots-clés : Codage automatique, Machine Learning, Terminologies médicales, Traitement automatique du langage.

1 Introduction

L'endoscopie digestive est un examen permettant de visualiser l'intérieur des organes du tube digestif et des voies bilio-pancréatiques, à l'aide d'un câble souple équipé d'une caméra ou d'un échographe, introduit par voie haute (bouche) ou par voie basse (anus). L'endoscopie peut être diagnostique (non interventionnelle), quand elle sert à établir un diagnostic, ou thérapeutique (interventionnelle), quand elle permet de traiter une maladie ou une lésion.

Dès lors qu'un geste d'endoscopie est réalisé, un compte rendu d'endoscopie est rédigé. Les comptes rendus d'endoscopie décrivent le motif de l'examen, le déroulement de celui-ci en détaillant les organes visualisés.

A la suite de l'examen endoscopique, les opérateurs de codage (médecin, technicien d'information médicale, secrétaire) doivent faire correspondre l'acte réalisé tel que décrit dans le compte rendu à un acte codé à l'aide de la Classification Commune des Actes Médicaux (CCAM). Le codage de l'acte est réalisé à des fins de standardisation du recueil de l'information médicale. L'information ainsi recueillie sert à la facturation mais également à des fins épidémiologique et de recherche.

La CCAM est une classification hiérarchique des actes médicaux. Elle ordonne ceux-ci selon des axes anatomiques et selon leur caractère diagnostique ou thérapeutique. A chaque acte est associé un code composé de quatre premiers caractères signifiants juxtaposés suivis de trois chiffres servant de compteur. Les deux premiers caractères déterminent la topographie, le troisième l'action et le quatrième la technique employée. Ainsi l'endoscopie recto-sigmoïdienne sera codée HJQE001, le H signifiant que l'acte sera réalisé sur l'appareil digestif, le J signifiant qu'il sera réalisé sur le rectum, le Q décrivant un acte d'examen et/ou d'enregistrement, et le E la voie endoscopique de l'acte.

Le codage des actes est une activité qui requiert une bonne connaissance de la CCAM, ainsi que des connaissances médicales . Cette double compétence requise est source d'erreurs et d'oublis (Rector, 1999). De plus, le montant de la prestation de soin facturé dépendant directement ou indirectement de ces codes, ces erreurs peuvent entraîner un écart de recette important pour l'établissement de soin. La proposition de codes CCAM à partir du contenu du compte rendu d'endoscopie permettrait de rendre plus reproductible cette tâche de codage.

Les méthodes d'apprentissage permettent à partir d'un volume conséquent de données d'élaborer des heuristiques pour résoudre une tâche. Leur utilisation pourrait permettre d'assigner un code CCAM en fonction du contenu du compte rendu d'endoscopie (Pakhomov *et al.*, 2006).

L'objectif est ici de développer une méthode d'apprentissage supervisé pour assigner automatiquement un code CCAM pour les endoscopies digestives diagnostiques à partir d'un compte rendu.

Après avoir présenté le corpus de texte sur lequel se réalisera la tâche d'apprentissage, les modalités de choix de l'algorithme, les performances de la solution retenue seront présentées et discutées.

2 Matériel et méthodes

Cette étude a été faite avec l'accord du chef de service de gastro-entérologie dans le cadre réglementaire de la réutilisation des données pour l'unité de soins.

2.1 Matériel

La CCAM a été utilisée pour identifier l'ensemble des actes d'endoscopie digestives non interventionnelles réalisables en ambulatoire. Cette sélection a été réalisée avec un gastro-entérologue. Ces actes sont décrits en Table 1. Les comptes rendus d'endoscopie réalisées au sein du service de Gastro-Entérologie de l'hôpital Avicenne entre 2015 et 2016, et correspondant à ces actes ont ensuite été récupérés. Ils n'ont pas été anonymisés, l'étude restant dans le périmètre de l'unité de soins, autorisant le traitement de données non anonymisées. Ces compte-rendus sont rédigés en texte libre par plus de dix médecins séniors différents. Ils suivent généralement un modèle qui précise les grandes catégories du compte rendu : Motif, Déroulement de l'examen, Conclusion. Les actes CCAM codés pour chaque compte rendu par ces mêmes médecins ont été également extraits du système de recueil du codage des actes. Le responsable du service de gastro-entérologie a également décrit de façon non ambigue les actes CCAM retenus en citant pour chaque acte les organes visualisés (Table 1).

Code CCAM	Libellé CCAM	Organes visualisés
HEQE002	Endoscopie œso-gastro-duodénale	Oesophage, estomac et duodénum
HHQE002	Coloscopie totale avec franchissement de l'orifice iléocolique	Anus, rectum, côlon sigmoïde, côlon descendant, transverse et ascendant, caecum, début de l'iléon
HHQE005	Coloscopie totale avec visualisation du bas-fond caecal, sans franchissement de l'orifice iléocolique	Anus, rectum, côlon sigmoïde, côlon descendant, transverse et ascendant, caecum (bas-fond caecal)
HJQE001	Rectosigmoïdoscopie	Anus, rectum et éventuellement côlon sigmoïde
HHQE004	Coloscopie partielle au-delà du sigmoïde	Anus, rectum, côlon sigmoïde et côlon descendant, éventuellement côlon transverse et ascendant

TABLE *1 – Description des actes CCAM retenus*

2.2 Méthodologie

2.2.1 Elaboration du "Gold Standard"

Afin de réaliser l'apprentissage supervisé et son évaluation, il est nécessaire de fournir une référence, c'est à dire un code CCAM correct pour chaque compte rendu. Cette liste de codes constituera les labels du gold standard. Pour constituer cette liste, le postulat a été de prendre en compte initialement les codes CCAM choisis par les médecins. Afin de contrôler leur qualité, une analyse manuelle par le médecin d'information médicale (DIM) des codes associés à 100 compte-rendus tirés aléatoirement a cependant révélé des fautes de codage récurrentes.

Pour y remédier, un algorithme a été développé. Il signale automatiquement les situations douteuses qui correspondent soit à l'absence dans le compte rendu de termes anatomiques attendus, ou à l'existence de termes anatomiques inattendus, en comparaison avec la description fournie par le gastro-entérologue. Les codes douteux ont été ensuite été corrigés par le DIM. Le label "Autre" a été attribué aux comptes rendus corrigés qui ne font pas partie du périmètre des endoscopies diagnostiques du tube digestif. Afin d'équilibrer les populations des différentes classes, la classe "Autre" a été repeuplée avec des comptes rendus d'endoscopie des voies biliaires ou des endoscopies interventionnelles. La liste comprenant ces codes corrigés, et les codes qui n'ont pas été signalés comme douteux constitue le gold standard.

2.2.2 Pré-traitement des comptes rendus

Avant de pouvoir exploiter les comptes rendus d'endoscopie, ceux-ci ont été pré-traités :

— Les accents et les caractères spéciaux ont été supprimés de façon automatique,
— Les acronymes (25) ont été répérés lors de la relecture aléatoire de 100 comptes rendus et remplacés par leur signification en toutes lettres.
— Certaines parties du compte rendu ont été supprimées afin que l'algorithme d'apprentissage n'identifie pas de mots discriminants dans des zones non fiables du compte rendus (par exemple en-tête : "Compte rendu de gastroscopie" utilisé pour un compte rendu de coloscopie). Ces suppressions ont été faites de façon automatatique.

2.2.3 Classification des comptes rendus par apprentissage

Le problème d'apprentissage est celui d'une classification multi-classe (cinq classes représentant un code CCAM, présentés en Table 1, et une classe "Autre"). Les comptes rendus ont chacun un label correspondant à son code CCAM ou à la classe "Autre" dans le gold standard.

2.2.3.1 Vectorisation

Les comptes rendus en texte brut ont été convertis en matrice contenant les mots et les bigrammes de chaque texte et leur "Text Frequency - Inverse Document Frequency" (TF-IDF), qui permet d'évaluer l'importance de chaque mot (ou bigramme) dans un document, relativement au corpus complet.

Les mots présents dans moins de trois documents ont été ignorés, considérés comme trop rares. Cela permet aussi d'éliminer des fautes d'orthographe rares. Les mots présents dans plus de 90 % des textes ont également été ignorés.

Les algorithmes de classification ont été testés avec les unigrammes seuls (mot simple) et avec unigrammes et bigrammes.

2.2.3.2 Evaluation de quatre méthodes de classification

La librairie *scikit-learn* (Pedregosa *et al.*, 2011) de Python fournit divers algortithmes de classification. Le "meilleur" algorithme a été choisi en comparant quatre modèles : *RandomForestClassifier*, *LinearSVC*, *MultinomialNB* et *LogisticRegression*. L'évaluation a été faite avec la macro-moyenne des précisions et rappels de chaque classe, puis la F-mesure par validation croisée à cinq plis.

L'algorithme avec la meilleure F-mesure choisi a été entraîné sur deux tiers aléatoires du corpus, et testé sur le tiers restant. Pour chaque classe, la précision, le rappel, et la F-mesure ont été calculés et les mots les plus corrélés à chaque classe ont été relevés.

Cet algorithme peut ensuite, en parcourant un nouveau compte rendu, lui attribuer un code.

3 Résultats

3.1 Corpus de comptes rendus et "Gold Standard"

L'algorithme développé pour détecter les erreurs de codages a permis d'identifier 202 codages douteux. Après leur vérification par le DIM, 115 ont été effectivement corrigés. Parmi les erreurs fréquentes de codage, on retrouve par exemple des endoscopies interventionnelles codées comme non interventionnelles, ou des confusions entre codes aux libellés proches. La Table 2 présente les taux d'alerte et de recodage pour chaque classe. La Table 3 présente les effectifs totaux de chaque classe après cette étape.

Code CCAM	Nombre de CR	Nombre de CR douteux	Pourcentage de CR douteux par classe	Nombre de CR à recoder	Pourcentage de CR à recoder par classe
HEQE002	1000	32	3.2 %	16	1.6 %
HJQE001	165	117	70.9 %	50	30.3 %
HHQE002	187	34	18.2 %	32	17.1 %
HHQE005	149	16	10.7 %	12	8.0 %
HHQE004	20	3	15 %	1	5 %
Total	1521	202	13.1 %	113	7.34 %

TABLE *2 – Nombre de comptes rendus signalés comme présentant un code douteux, et nombre de codes effectivement faux pour chaque classe*

Code CCAM	Nombre de comptes rendus
HEQE002	985
HJQE001	118
HHQE002	172
HHQE005	162
HHQE004	70
"Autre"	32
Total	1539

TABLE *3 – Population des différentes classes après corrections*

3.2 Classification des comptes rendus

3.2.1 Sélection du modèle

Le modèle Linear SVC, avec unigrammes seuls, présente la meilleure F-Mesure moyenne (Table 4). C'est le modèle choisi pour la suite.

Modèle	F-mesure moyenne (unigrammes et bigrammes)	F-mesure moyenne (unigrammes seuls)
LinearSVC	0.822	0.833
Logistic Regression	0.668	0.676
Multinomial Naive Bayes	0.552	0.537
Random Forest	0.531	0.552

TABLE *4 – F-mesure moyenne pour les modèles de classification testés par validation croisée à 5 plis*

3.2.2 Evaluation du modèle LinearSVC

Un tiers du corpus (508 comptes rendus), tiré au sort, a été reservé au test.
Les métriques d'évaluation de la méthode sont présentées en Table 5.

Classe (code)	Précision	Rappel	F-mesure	Effectif
HEQE002	0.98	1	0.99	325
HHQE002	0.90	0.90	0.90	52
HHQE005	0.87	0.92	0.90	51
HHQE004	0.82	0.62	0.71	29
HJQE001	0.85	0.91	0.88	44
Autre	1	0.43	0.60	7
Moyenne pondérée/total	0.94	0.94	0.94	508

TABLE *5 – Métriques pour chaque classe, pour 508 comptes rendus testés avec la méthode Linear SVC*

Pour chaque classe, les mots les plus corrélés sont :
— HEQE002 : oesophage, estomac, normal, gastrique, pylore
— HHQE002 : ileon, ileocoloscopie, derniere, ileale, sur
— HHQE005 : totale, caecum, coloscopie, caecal, fond
— HHQE004 : gauche, colique, jusqu, colon, incomplete
— HJQE001 : marge, anale, normacol, rectosigmoidoscopie, charniere

4 Discussion

L'algorithme choisi donne une précision et un rappel moyens de 0,94, avec de très bons résultats pour les trois classes les plus représentées. Ce sont les actes les plus fréquemment réalisés, mais on remarque en Table 2 que les coloscopies totales (HHQE002 et HHQE005) sont régulièrement mal codées.

On retrouve dans la littérature des résultats inférieurs pour ce type de tâche par apprentissage supervisé (extraction de diagnostics) avec une précision moindre ou un rappel moindre selon les méthodes utilisées (Wang *et al.*, 2012). Cela peut être expliqué par le nombre faible d'actes à coder dans notre cas.

Les mots saillants rapportent des noms d'actes (coloscopie, iléocoloscopie, rectosigmoïdoscopie) cohérents, ainsi que des termes anatomiques décrivant logiquement les organes visualisés. Ces résultats sont comparables avec ceux obtenus en extraction supervisée de diagnostics, qui révèlent des termes anatomiques et des termes correspondant au diagnostic recherché : cancer, adénocarcinome, tumeur par exemple pour un cancer (Wang *et al.*, 2012).

Le corpus étudié comporte 1539 comptes rendus. Ce nombre est relativement faible en comparaison avec le volume du corpus d'autres travaux de classification ou de codage de

comptes rendus médicaux, qui varient entre 4500 pour les études restreintes à un domaine précis (Saad *et al.*, 2006) et 2 millions pour des classification très générales (Pakhomov *et al.*, 2006). Cependant, le nombre de codes cibles (5 + 1) est suffisamment réduit pour permettre des résultats corrects.

L'algorithme de détection des codes douteux a permis de révéler une qualité moyenne du gold standard. De plus, en l'absence d'une vérification complète (et manuelle) des codes correspondant aux comptes rendus du corpus, il est impossible d'évaluer cet algorithme. Nous nous retrouvons dans le cas d'une classification utilisant des labels bruités (Agarwal *et al.*, 2016).

On remarque que les termes saillants sont majoritairement des termes anatomiques, il aurait été possible d'élaborer une méthode de repérage de ces termes par pur traitement du langage, avec par exemple une détection de la négation. Des méthodes similaires sont utilisées avec des résultats supérieurs à 0.94 en rappel et précision dans le cadre de l'extraction de diagnostics dans les comptes rendus de radiologie pulmonaire (Friedlin & McDonald, 2006).

Une perspective serait de deviner de façon automatique le code CCAM "lettre par lettre" d'un acte en utilisant la signifiance de chaque lettre. La recherche de l'organe jusqu'auquel on progresse lors de l'endoscopie permet de trouver la deuxième lettre du code (E pour œsophage, H pour le côlon par exemple).

5 Conclusion

La méthode développée ici permet de coder automatique des comptes rendus semi-structurés d'endoscopie digestive, avec de bons résultats pour le périmètre choisi, et de détecter de potentielles erreurs de codage.

6 Remerciements

Nous remercions le Professeur Robert Benamouzig, chef du service de Gastro-Entérologie de l'hôpital Avicenne, d'avoir autorisé cette étude sur les comptes rendus d'endoscopie de son service.

Références

AGARWAL V., PODCHIYSKA T., BANDA J. M., GOEL V., LEUNG T. I., MINTY E. P., SWEENEY T. E., GYANG E. & SHAH N. H. (2016). Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association : JAMIA*, **23**(6), 1166–1173.

FRIEDLIN J. & MCDONALD C. J. (2006). A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, p. 269–273.

PAKHOMOV S. V. S., BUNTROCK J. D. & CHUTE C. G. (2006). Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. *Journal of the American Medical Informatics Association*, **13**(5), 516–525.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RECTOR A. L. (1999). Clinical terminology : why is it so hard ? *Methods of Information in Medicine*, **38**(4-5), 239–252.

SAAD F. H., IGLESIA B. D. L. & BELL G. D. (2006). Comparison of Documents Classification Techniques to Classify Medical Reports. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, p. 285–291 : Springer, Berlin, Heidelberg.

WANG Z., SHAH A. D., TATE A. R., DENAXAS S., SHAWE-TAYLOR J. & HEMINGWAY H. (2012). Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. *PLoS ONE*, **7**(1).