Classifier Chains pour le codage LOINC

T. MICHEL–PICQUE[1,3], S. BRINGAY[1,2], P. PONCELET[1], N. PATEL[2,3], G. MAYORAL[3]

[1] LIRMM UMR 5506, University of Montpellier, CNRS, Montpellier, France
[2] AMIS, Paul-Valéry University, Montpellier, France
[3] Onaos, Montpellier, France

tmichelpicque@lirmm.fr

Résumé

Objectif : *Cet article présente une étude sur le codage de données réelles issues de laboratoires français vers la terminologie LOINC.* ***Méthodes :*** *Nous présentons une comparaison de trois approches pour le codage LOINC. Ces approches incluent à la fois une approche de modèle linguistique de l'état de l'art ainsi qu'un classifier chains.* ***Résultats :*** *Notre étude démontre que nous améliorons avec succès la performance de labaseline en utilisant le classifier chains et que nous nous comparons efficacement aux modèles linguistiques de l'état de l'art.* ***Conclusions :*** *Bien que notre approche rencontre des défis de reproductibilité et présente des perspectives d'optimisations et de tests futurs sur des ensembles de données, elle s'avère néanmoins efficace et économique.*

Mots-clés

Codage LOINC, Interopérabilité, Standardisation des Données Cliniques, Modèle linguistique

Abstract

Purpose : *This article presents a study on the coding of real data from French laboratories into the LOINC terminology.* ***Methods :*** *We present a comparison of three approaches for LOINC coding. These approaches include both a state-of-the-art language model approach and a classifier chains approach.* ***Results :*** *Our study demonstrates that we successfully improve the performance of the baseline using the classifier chains approach and compete effectively with state-of-the-art language models.* ***Conclusions :*** *Our approach proves to be efficient and cost-effective despite reproducibility challenges with perspectives for future optimizations and dataset testing.*

Keywords

LOINC Coding, Interoperability, Clinical Data Standardization, Language model

1 Introduction

LOINC [1] [8] (Logical Observation Identifiers Names and Codes) est une terminologie de référence internationale employée depuis la fin des années 1990 pour identifier et étiqueter les observations cliniques et les résultats de tests médicaux. Elle établit un système de codage standardisé pour les tests de laboratoire, les mesures cliniques et diverses observations de santé. Chaque code LOINC comprend une combinaison de chiffres et de caractères alphanumériques avec une équivalence au format textuel, délimitant précisément la nature de l'analyse ou du résultat. LOINC a été largement adopté dans les établissements de santé du monde etier, facilitant l'échange de données cliniques entre les systèmes d'information de santé et assurant une interopérabilité efficace. En France, cette terminologie a été adoptée officiellement pour la biologie médicale depuis 2016. Le ministère du travail, de la santé et des solidarités à travers l'ANS (l'Agence du Numérique en Santé) a élaboré dans le cadre du Ségur du Numérique en 2016 une feuille de route pour coder l'ensemble des laboratoires de biologie médicale [2].

Cet article décrit une procédure de codage LOINC en français, visant à établir un codage automatisé entre les examens de laboratoire catalogués localement et leurs codes LOINC correspondants. L'article apporte une double contribution. Grâce à des expériences approfondies visant à standardiser les données de laboratoire françaises selon le format LOINC, nous obtenons deux résultats clés : 1) nous démontrons l'efficacité de l'état de l'art dans le codage de codes LOINC basé sur des modèles linguistiques, et 2) nous proposons une alternative peu couteuse qui, basée sur la méthode de classifier chains [7], améliore sensiblement les résultats. Le reste de l'article est organisé comme suit : La section 2 présente l'état de l'art dans le domaine du codage LOINC puis 3 décrit les méthodes mises en œuvre, tandis que la section 4 fournit des détails sur les expériences menées et présente les résultats. Enfin, nous concluons en discutant de nos travaux futurs.

2 Travaux existants

Dans le contexte du codage de code LOINC, des recherches précédentes ont démontré l'utilité et la pertinence d'outils de codage manuels ou semi-manuels, souvent gérés par des experts médicaux [6]. Cependant, en raison de la nature chronophage de la tâche et de l'impraticabilité de la délé-

1. https ://loinc.org

2. https ://industriels.esante.gouv.fr/segur-numerique-sante/vague-1/dispositif-loinc-couloir-biologie-medicale

Code client	Libellé local	Synonymes	NABM	Échelle
AC	AC ANTINUCLEAIRE	Auto-anticorps antinucléaires	324	Numérique
Technique	**Code chapitre**	**Système**	**Tube échantillon**	**Unités**
Frottis	Calcul	Sérum	M4RT	g/l

TABLE 1 – Exemples illustratifs d'une entrée x_i

guer à des individus moins formés que des experts médicaux, cette méthode n'est actuellement pas l'approche privilégiée. Avec l'adoption généralisée de l'apprentissage automatique, de nouvelles avancées dans ce domaine ont déjà été réalisées. Par exemple, Kelly et al. [2] ont développé des méthodologies de codage innovantes pour vectoriser les données de laboratoire en texte libre et ont évalué la performance de modèles d'apprentissage automatique tels que la régression logistique, les forêts aléatoires, et les K-plus proches voisins pour le codage LOINC. Tu et al. [9] tirent parti de plongements contextuels à partir de modèles T5 pré-entraînés et proposent une stratégie de fine-tuning en deux étapes basée sur l'apprentissage contrastif. Ai et al. [1] ont modifié la structure des données en concaténant les caractéristiques d'un coté et les cibles de l'autre pour effectuer une classification de phrases avec des réseaux siamois. Bien que ces méthodes soient très efficaces, celles basées sur les modèles linguistiques présentent des défis en termes d'interprétation et soulèvent des considérations éthiques et de biais en raison de leur potentiel à hériter des biais sociétaux présents dans les données, menant à des résultats biaisés ou injustes. De plus, l'optimisation de ces modèles pour des tâches spécifiques peut être coûteux et peut nécessiter un grand volume de données qui est compliqué à obtenir car les données proviennent de multiples laboratoires de biologie privés indépendants les uns des autres. Enfin, ces modèles négligent souvent les interconnexions entre les différentes parties du code LOINC. Dans cet article, s'inspirant des travaux de Read et al. [7] sur les classifiers chains, nous n'abordons pas le codage comme une simple tâche de classification multi-étiquettes. Au contraire, nous considérons les corrélations entre les composants de l'étiquette, proposant une nouvelle méthode de chains classifiers qui modélise ces corrélations de composants.

3 Méthodes

3.1 Description de la tâche

Étant donné un catalogue d'examens d'un laboratoire français, le codage consiste à prédire le code LOINC c'est-à-dire un code numérique tel que '42254-3', unique pour chaque examen de biologie médicale en se basant sur les informations présentes pour chaque examen dans ce catalogue. Comme illustré dans la Figure 1, ce code a un équivalent textuel connu sous le nom de 'LOINC complet' (par exemple, 'Noyau anticorps [Présence/Seuil] ; Serum ; Qualitatif') qui peut être décomposé en attributs, tels que le Milieu (System) (par exemple, 'Sérum'), Composant (par exemple, 'Noyau anticorps'), Échelle (par exemple, 'Qualitatif'), etc. Considérant un ensemble de données D comprenant des résultats d'examen de laboratoires français, chaque résultat d'examen x_i est représenté par un vecteur de caractéristiques utilisé pour la prédiction. Un exemple illustratif d'une telle entrée est fourni dans le Tableau 1. La cible de prédiction pour chaque examen, notée y_i, est un vecteur défini comme $y_i = S_i, P_i, Sc_i, Lac_i, Lcomp_i, Cl_i$, où S_i représente le Milieu, P_i la Propriété, Sc_i l'échelle, Lac_i le sous Composant (ou Analyte core), $Lcomp_i$ le Composant, et Cl_i le LOINC complet.

Comme illustré dans la Figure 1, la relation hiérarchique entre Cl, $Lcomp$, et Lac peut être représentée à travers une hiérarchie d'inclusion, où $Lac \subset Lcomp \subset Cl$. Cela indique que Lac est inclus dans $Lcomp$ et $Lcomp$ est inclus dans Cl. De plus, il convient de noter que S, Sc, et P sont également inclus dans Cl, soulignant davantage la structure étendue et imbriquée de ces relations au sein de la hiérarchie. L'objectif est de développer un modèle de prédiction $f : X \rightarrow Y$ pour coder avec précision les vecteurs de caractéristiques d'entrée X aux vecteurs cibles Y, en tenant compte de la structure hiérarchique des composants LOINC.

3.2 Modèles

Ci-dessous, nous offrons un aperçu complet des trois modèles que nous avons testés.

- **Baseline** : Afin de traiter le codage comme un problème multi-classes, nous exploitons la structure tabulaire des caractéristiques d'entrée pour prédire la cible finale Cl_i.
- **Text-based** : Nous traitons le codage comme un problème de classification de phrases en prédisant la classe finale Cl_i à partir de la concaténation des caractéristiques d'entrée.
- **Classifier chains** : Au lieu de prédire uniquement la cible finale du codage Cl_i, nous commençons par des cibles intermédiaires (c'est-à-dire des attributs de Cl_i) et les incorporons comme caractéristiques pour prédire la cible suivante, menant finalement à la prédiction de la cible finale Cl_i.

4 Expériences et Résultats

4.1 Expériences

Les expériences ont été menées sur un ensemble de données réelles collectées auprès de 99 laboratoires français, comprenant 162,678 entrées pour 11,216 codes LOINC uniques. Nous avons affiné cette distribution en ne conservant un code que s'il apparaît au moins 10 fois, ce qui donne un ensemble de données final de 139,235 entrées et 2,463 codes LOINC. Nous avons utilisé différentes implémentations pour chaque méthode.

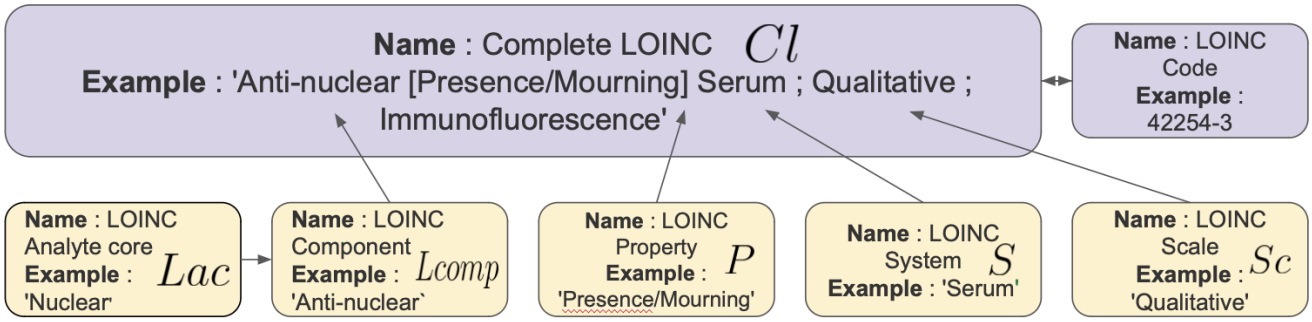

FIGURE 1 – Organisation hiérarchique des attributs LOINC

— **Baseline** : Nous avons encodé les caractéristiques d'entrées avec un `OneHotEncoder` afin de convertir ces variables catégorielles en une représentation binaire. Chaque catégorie unique dans une variable est transformée en une colonne binaire distincte. La cible est créée avec un `LabelEncoder` afin de convertir les variables catégorielles en valeurs numériques. Chaque catégorie est représentée par un entier unique. Avec ce prétraitement, nous avons pu entraîner des modèles `SGD` et `XGBoost` pour prédire la cible finale Cl_i en abordant le problème comme une classification multi-classes. Nous avons choisi ces deux modèles car `SGD` s'est révélé efficace et rapide sur des jeux de données en grande dimension, ce qui est notre cas avec notre `OneHotEncoder`, et `XGBoost` est reconnu pour être un modèle d'apprentissage automatique classique, robuste et très performant. Les hyperparamètres par défaut ont été utilisés pour les deux modèles. Les bibliothèques utilisées pour ces modèles sont `scikit-learn` pour `SGD` et `xgboost` pour `XGBoost`.

— **Text-Based** : Nous avons commencé par concaténer et tokeniser tous les composants textuels des entrées de laboratoire en utilisant comme séparateur `';'`. Pour la cible, nous utilisons le LOINC complet Cl. Ensuite, nous avons utilisé `CamemBERT` [5], un modèle linguistique français, et `DrBERT` [3], un modèle linguistique biomédical français. Ce sont tous deux des modèles de l'état de l'art basés sur `RoBERTa` [4]. Nous les avons entraînés pour une tâche de classification pendant 20 epochs, en utilisant un batch de 64 et l'optimiseur `AdamW`.

— **Classifier chains** : Initialement, nous avons encodé chaque cible avec une instance différente de `LabelEncoder` et avons initialisé un classifier pour chaque étape de notre algorithme. Nous avons ensuite encodé nos caractéristiques avec un `OneHotEncoder`, répétant l'opération chaque fois que nous avons incorporé la cible précédente (prédit dans notre entrée avec le même prétraitement). Nous utilisons les mêmes classifiers et les mêmes hyperparamètres que pour la baseline pour faciliter une comparaison de nos approches.

Pour tous nos entraînements, nous avons utilisé une machine équipée de 45 GB de RAM, 16 cœurs CPU, et un GPU Tesla V100S avec 12 GB de VRAM.

4.2 Résultats

Pour évaluer la performance de nos approches, nous avons utilisé la précision, le rappel et le score F1 comme métriques. Toutes les métriques sont pondérées pour tenir compte de l'hétérogénéité de la distribution des données. Pour cela, nous avons utilisé le paramètre `average='weighted'` lors de notre calcul de performance avec la librairie `scikit-learn`, ce qui permet de réaliser une moyenne pondérée de nos métriques en fonction de notre distribution. Les résultats sont présentés dans le Tableau 2.

Les résultats de notre comparaison révèlent plusieurs constatations. Toutes les méthodes montrent de bons résultats, avec le score F1 le plus bas enregistré pour la baseline `XGBoost` à $0,80$. Les performances des deux modèles basés sur le texte sont très similaires. Cela suggère que la spécialisation de CamemBERT pour la langue française et celle de DrBERT pour les données médicales est suffisamment marquée pour entraîner une différence notable. La meilleure méthode est l'approche classifier chains utilisant un `SGD`, atteignant un score F1 de $0,87$ et qui surpasse les autres dans les deux autres métriques (précision de $0,88$ et rappel de $0,87$).

5 Limites de l'étude

Dans ce travail, les expérimentations restent préliminaires. Du fait de la structuration de LOINC en six dimensions, il est peu pertinent de faire une classification directe et les résultats des expériences (Baseline SGD et Baseline XGboost) étaient attendues. Même si on contaste une amélioration avec les expériences (Text-based CamemBERT et Text-based DrBERT), celle-ci reste limitée. Les expériences sur le Classifier chains améliorent les résultats mais elles doivent encore être complétées pour faire le pendant des expériences précédentes, où l'encodage du texte par un modèle linguistique doit être utilisé dans le contexte d'une

Méthode	Précision (Pondérée) (%)	Rappel (Pondéré) (%)	Score F1 (Pondéré) (%)
Baseline SGD	0,85	0,83	0,84
Baseline XGboost	0,85	0,75	0,80
Text-based CamemBERT	0,84	0,80	0,82
Text-based DrBERT	0,81	0,81	0,81
Classifier chains SGD	**0,88** (+0,03)	**0,87** (+0,04)	**0,87** (0,03)
Classifier chains XGBoost	0,81 (-0,04)	0,82 (+0,07)	0,82 (+0,02)

TABLE 2 – Comparaison des résultats de nos différentes approches

détermination séparée des six attributs.

Une deuxième limitation importante est la non prise en compte de la distribution déséquilibrée des codes LOINC. Avoir sélectionné les codes ayant au moins 10 occurrences a rendu notre tâche envisageable. Toutefois, même dans ces conditions, la fréquence des codes impacte les performances du système. Une analyse de la nature des valeurs des différents champs serait aussi importante car certains champs ont probablement peu de valeurs différentes alors que d'autres beaucoup.

6 Discussions et Conclusions

Dans ce travail, nous avons mené des expériences approfondies sur le codage LOINC en français. Nous avons comparé des approches basiques d'apprentissage automatique, une approche de l'état de l'art basée sur les modèles linguistiques, et une méthode novatrice utilisant un classifier chains, qui prend en compte les liens de dépendance entre les différentes parties du LOINC. L'approche classifier chains utilisant un `SGD` s'est révélée être la plus efficace, atteignant un score F1 de $0,87$. Les premiers résultats obtenus à partir de données réelles sont prometteurs, signalant la nécessité d'expériences supplémentaires et d'une exploration plus approfondie. Cependant, comparer les performances avec d'autres études peut être difficile étant donné la diversité dans le nombre de codes LOINC couverts. Il est également concevable que les modèles linguistiques pourraient donner de meilleures performances avec des optimisations et des ajustements supplémentaires. Enfin, notre modèle classifier chains pourrait faire face à des défis notamment à un changement de la distribution sur un autre ensemble de données étant donné qu'il s'agit de catalogues d'examens spécifiques à chaque laboratoire français, un examen pouvant être représenté de deux façons différentes mais équivalentes. Comme prochaine étape, nous prévoyons de tester cette approche sur divers ensembles de données, y compris le jeu de données MIMIC pour augmenter la robustesse et l'interopérabilité entre français et anglais.

7 Remerciements

Cet article est basé sur des travaux soutenus par l'ANRT (Association nationale de la recherche et de la technologie) avec une bourse CIFRE accordée à MICHEL-PICQUE Théodore.

Références

[1] Dige Ai, Yu He, Shenghai Jin, Xuemin Liu, Nianyi Sun, Guangku Tian, and Zhiqiang Zhang. A novel deep learning model for automated mapping of chinese laboratory test terminologies to logical observation identifiers names and codes (LOINC). *Available at SSRN 4092365*, 2022.

[2] Jonathan Kelly, Chen Wang, Jianyi Zhang, Spandan Das, Anna Ren, and Pradnya Warnekar. Automated mapping of real-world oncology laboratory data to LOINC. In *AMIA Annual Symposium Proceedings*, volume 2021, page 611. American Medical Informatics Association, 2021.

[3] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. DrBERT : A robust pretrained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04, 2023.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

[5] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[6] Jean Noël Nikiema, Romain Griffier, Vianney Jouhet, and Fleur Mougin. Aligning an interface terminology to the logical observation identifiers names and codes (LOINC). *JAMIA open*, 4(2) :ooab035, 2021.

[7] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine learning*, 85 :333–359, 2011.

[8] Roberto A Rocha and Stanley M Huff. Coupling vocabularies and data structures : lessons from LOINC. In *Proceedings of the AMIA Annual Fall Symposium*, page 90. American Medical Informatics Association, 1996.

[9] Tao Tu, Eric Loreaux, Emma Chesley, Adam D Lelkes, Paul Gamble, Mathias Bellaiche, Martin Seneviratne,

and Ming-Jun Chen. Automated LOINC standardization using pre-trained large language models. In *Machine Learning for Health*, pages 343–355. PMLR, 2022.