Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-107

Development of an Automatic Coding System for Digestive Endoscopies

Iris TERNOIS[a,b,1], Jean-Baptiste ESCUDIE[a], Robert BENAMOUZIG[a] and Catherine DUCLOS[a,b]
[a] *Hopital Avicenne, APHP, Bobigny, France*
[b] *LIMICS, Universite Paris 13, INSERM UMRS 1142, Bobigny, France*

Abstract. Digestive endoscopies, along with all medical procedures in France are coded with the CCAM. This task is done by the physicians, is time-consuming and requires a good knowledge of the terminology besides a medical knowledge. This method offers an automatic coding of endoscopic procedures from free-text reports. Thanks to a supervised learning method, the reports are coded with an average precision and recall of 0.92 on a 1639 texts corpus.

Keywords. Automatic Coding, Machine Learning, Medical Terminology, Natural Language Processing

1. Introduction

Digestive endoscopy is a medical exam allowing to view the interior of the digestive system by means of a flexible cable fitted with a camera, inserted through the mouth or the anus. The endoscopy can be diagnostic or therapeutic depending on the aim of the physician (diagnose or treat a disease or lesion).

After an endoscopy, the physician writes a report describing the motive, the process and the organs viewed. Then coding operators have to correlate the procedure described in the report to a code from the French classification of procedures called CCAM. This code will be used for billing and research.

The CCAM is a hierarchic classification that organizes procedures by an anatomical axis and according to their feature (diagnostic or therapeutic). Each procedure is linked to a four letters code followed by three numbers. The two first letters identify the topography of the procedure (e.g. HJ for rectum), the third the action (Q for examination or record) and the fourth the technique used (E for endoscopic approach, thus HJQE stands for recto sigmoid endoscopy).

Coding procedures requires a strong knowledge of the CCAM, as well as medical knowledge. This ambivalence is a source of error and omissions [1].

Machine learning methods enable to elaborate heuristics from large volume of data in order to solve a task. Their use could allow the assignation of a CCAM code depending on the content of the medical report [2].

The aim is to develop a supervised learning method that assigns automatically a code to a diagnostic digestive endoscopy report.

[1] Corresponding Author, Iris Ternois, LIMICS, Université Paris 13, 74 rue Marcel Cachin, 93017 Bobigny Cedex, France ; E-mail: iris.ternois@gmail.com.

After introducing the corpus of texts on which the learning task is carried out, the selection of the algorithm and its evaluation will be presented and discussed.

2. Method

2.1. Material

This study has been carried out with the consent of the head of the gastro-enterology department in the hospital (Avicenne in Bobigny, France). 3586 endoscopy reports written between 2015 and 2016 have been collected. They are free-text and follow a template specifying the categories: Motive, Progress and Conclusion. Their associated CCAM codes have been extracted. The CCAM has been used in order to identify all the diagnostic digestive endoscopies ambulatory achievable. This selection was made with the help of the gastro-enterologist. He also specified the anatomical bounds of each procedure.

2.2. Elaboration of the Gold Standard

A gold standard has been created by using the initial CCAM codes and cleaning it using a semi-automatic method. It flags a report when the associated code does not match with the requirements of its description: such as a report coded like a complete colonoscopy but not containing a mention of the caecum. The label "Other" has been attributed to the reports not in the scope of the study (non-diagnostic digestive endoscopies). 118 reports have been added to the corpus to repopulate the class "Other", so that the algorithm will also be able to tell if a report corresponds to a diagnostic endoscopy or not.

2.3. Pre-processing of the Reports

Each report has been pre-processed to simplify the learning task. French accents and special characters have been removed. Acronyms regularly used (25) have been written in full. Some parts of the reports have been suppressed (the header for instance) so that the algorithm does not use unreliable zones of them.

2.4. Reports Classification

This learning method classifies the reports into six different classes: five different CCAM codes and a class "Other". Each report has a label from the gold standard.

2.5. Vectorization

Free-text reports have been converted into a matrix containing each unigram and bigram and their TF-IDF (Text Frequency-Inverse Document Frequency) score. The rare (present in less than three reports) and very frequent (present in more than 90 %) words have been ignored. The classification algorithms have been tested with unigrams only (words) and with unigrams and bigrams.

2.6. Evaluation of Four Methods

The best algorithm was chosen between four algorithms provided by the Python Scikit-learn library: *RandomForestClassifier*, *LinearSVC*, *MultinomialNB*, and *LogisticRegression*. The algorithm with the best macro average F1 score, obtained by 5-fold cross-validation, has been trained on two thirds of the corpus and tested on the remaining third. For each class, precision, recall and F1 score have been calculated and most correlated words identified. This algorithm will then be able to attribute a CCAM code to a new endoscopy report.

3. Results

3.1. Diagnostic Digestive Endoscopies

Diagnostic digestive endoscopies are represented by 21 CCAM codes, divided into three types (video-endoscopies, echo-endoscopies and video capsule endoscopy). The echo-endoscopies and video capsule exam are excluded from the study, whether they are too rare or obsolete in 2018. Amongst the 13 codes remaining, 5 codes are used correctly. They are described in Table 1.

3.2. Corpus and Gold Standard

The results are displayed in Table 1.

Table 1. Distribution of reports based on the codes

CCAM code	Initial number of reports	Number of reports after correction	CCAM denomination and procedure description
HEQE002	1000	985	Oeso-gastro-duodenal endoscopy: esophagus, stomach and duodenum exploration (orally)
HHQE002	187	172	Complete colonoscopy with crossing of the ileo-colic orifice: colon and beginning of ileum exploration
HHQE005	149	162	Complete colonoscopy with caecum visualization and no crossing of the ileo-colic orifice: endoscopy up to the caecum
HJQE001	165	118	Recto-sigmoid endoscopy: anal canal, rectum and sigmoid exploration
HHQE004	20	70	Partial colonoscopy: endoscopy up to the left, transverse or right colon.
Other	0	132 (14+118)	
Total	1521	1639	

3.3. Reports Classification

3.3.1. Algorithm Selection

The results are displayed in Table 2.

Table 2. Average f-score for each algorithm, after 5-fold cross-validation

Algorithm	Average f-score (unigrams and bigrams)	Average f-score (unigrams only)
Linear SVC	0.831	0.822
Logistic Regression	0.726	0.743
Multinomial Naïve Bayes	0.597	0.613
Random Forest	0.539	0.546

3.3.2. Linear SVC Evaluation

One random third of the corpus (541 reports) have been saved for the test. For each class, most correlated words are (they have been translated from French to English):

- HEQE002: esophagus, stomach, normal, gastric, pylorus
- HHQE002: ileum, ilecolonoscopy, last, ileal, over
- HHQE005: colonoscopy, right, caecal, bottom, colon
- HHQE004: left, colic, angle, colon, up to
- HJQE001: anal, recto sigmoid endoscopy, anus, junction, rectum
- Other: remove, jejunum, millimeter, tube, adenopathy

The metrics for each class are displayed in Table 3.

Table 3. Metrics for each class (CCAM code or "Other")

Class (code)	Precision	Recall	F-score	Population of the class
HEQE002	0.97	1	0.98	330
HJQE001	0.89	0.93	0.91	44
HHQE002	0.83	0.88	0.85	50
HHQE002	0.82	0.89	0.85	55
HHQE004	0.88	0.60	0.71	23
Other	0.81	0.57	0.67	37
Weighted average/Total	0.92	0.92	0.92	541

4. Discussion

The algorithm we used produces an average precision and recall of 0.92, and excellent results for the most represented classes. For instance, recto sigmoid endoscopies are coded with a precision of 0.89, whereas they are miscoded in 30 % of the cases when manually coded.

The literature exposes lower results for this type of task (e.g. Diagnoses extraction using supervised learning methods), with lower precision or lower recall depending on the method [4]. This may be explained by the limited number of codes in this study. The most relevant words return coherent procedures names (e.g. colonoscopy, recto sigmoid

endoscopy) as well as anatomical names logically describing the organs viewed. Those results are comparable to the ones obtained with a diagnoses extraction, that reveal anatomical names and diagnoses related words (for instance, cancer, adenocarcinoma, tumor, neoplasm for a cancer) [4].

The study is carried out on relatively small number of reports (1639) compared to other classification or automatic coding studies, for which volume of data ranges from 4500 texts for restricted perimeters [5] to 2 million texts for very general classifications [6]. However, the number of classes in our study is small enough to obtain good results.

The algorithm that detects possible wrong codes reveals the quite low quality of the initial coding. Furthermore, in the absence of an entire re-coding of the corpus by specialists, it is not possible to evaluate this algorithm. We find ourselves in a situation of supervised learning with noisy labels [6].

Most of the discriminating words are anatomical terms. They would be identified by a natural language processing method. Similar methods have been employed with satisfying results (precision and recall over 0.94) for the extraction of diagnoses in pulmonary radiology reports [7].

5. Conclusion

This method enables the automatic coding of semi-structured endoscopy reports, with satisfying results for our scope, and the detection of possible wrong codes.

References

[1] A.L. Rector, Clinical Terminology: why is it so hard?, *Methods of Information in Medicine* **38** (1999), 239-252.

[2] S.V.S. Pakhomov, J.D. Buntrock, C.G. Chute, Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques, *Journal of the American Medical Informatics Association* **13** (2006), 516-525.

[3] Z. Wang, A.D. Shah, A.R. Tate, S. Denaxas, J. Shawe-Taylor, H. Hemingway, Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning, *PLoS ONE* **7** (2012).

[4] F.H. Saad, B.D.L. Iglesia, G.D. Bell, Comparison of Documents Classification Techniques to Classify Medical Reports in *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin, 2006.

[5] V. Agarwal, T. Podchiyska, J.M. Banda, V. Goel, T.I. Leung, E.P. Minty, T.E. Sweeney, E. Gyang, N.H. Shah, Learning statistical models of phenotypes using noisy labeled training data, *Journal of the American Medical Informatics Association* **23** (2016), 1166-1173.

[6] J. Friedlin, C.J. MCDonald, A natural language processing system to extract and code concept relating to congestive heart failure from chest radiology reports, *AMIA Annual Symposium proceedings* (2006), 269-273.