Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220656

Determining the Set of Items to Include in Breast Operative Reports, Using Clustering Algorithms on Retrospective Data Extracted from Clinical DataWarehouse

Adrien BOUKOBZA[a,b,c,1], Maxime WACK[a,b,c], Antoine NEURAZ[a,b,c] , Daniela GEROMIN[d], Cécile BADOUAL[d], Anne-Sophie BATS[e,f], Anita BURGUN[a,b,c], Meriem KOUAL[f,g] and Rosy TSOPRA[a,b,c]

[a]*Université de Paris, Inserm, Sorbonne Université, Centre de Recherche des Cordeliers, F-75006 Paris, France*
[b]*HeKA, Inria Paris, France*
[c]*Department of Medical Informatics, Hôpital Européen Georges-Pompidou and Hôpital Necker - Enfants Malades, AP-HP, Paris, France*
[d]*Plateforme Centre de Ressources Biologiques et Tumorothèque, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*
[e]*INSERM UMR-S 1147, Université de Paris, Paris, France*
[f]*Chirurgie cancérologique gynécologique et du sein, Hôpital Européen Georges-Pompidou, AP-HP, Paris, France*
[g]*Université de Paris, Laboratoire INSERM 1124- Equipe 1 METATOX, Paris, France*

Abstract. Medical reports are key elements to guarantee the quality, and continuity of care but their quality remains an issue. Standardization and structuration of reports can increase their quality, but are usually based on expert opinions. Here, we hypothesize that a structured model of medical reports could be learnt using machine learning on retrospective medical reports extracted from clinical data warehouses (CDW). To investigate our hypothesis, we extracted breast cancer operative reports from our CDW. Each document was preprocessed and split into sentences. Clustering was performed using TFIDF, Paraphrase or Universal Sentence Encoder along with K-Means, DBSCAN, or Hierarchical clustering. The best couple was TFIDF/K-Means, providing a sentence coverage of 89 % on our dataset; and allowing to identify 7 main categories of items to include in breast cancer operative reports. These results are encouraging for a document preset creation task and should then be validated and implemented in real life.

Keywords. Machine learning, NLP, Clustering, Breast cancer

1. Introduction

Medical reports are key elements to guarantee the quality, safety, and continuity of care. However, their quality remains an issue, due to missing or inaccurate information [1]. A

[1] Corresponding Author, Adrien Boukobza, Université de Paris, Faculté de médecine, Paris, France; E-mail: adrien.boukobza@outlook.fr

few methods have been suggested to improve their quality, such as the extraction of patient data from electronic health record or the standardization of medical reports. Standardization of medical reports consists of determining a set of items to include in medical reports. It is usually addressed using guidelines from scientific societies or expert groups [2]. Nevertheless, standardization process comes with some limits: (i) it is a complex and tedious process (e.g., multiple experts, Delphi method); (ii) the set of items is too generic and not always adapted to the cause of hospitalization; (iii) the set of items is not adapted to local practices leading to a risk of non-adoption by physicians.

We hypothesize that a structured model of medical reports could be learnt using machine learning algorithms applied to retrospective medical reports from clinical data warehouses (CDW) [3]. Machine learning techniques, and especially clustering algorithms, have shown their value in extracting medical information from medical reports, alleviating time and resource consumption for this task [4–6].

Here, we propose to investigate our hypothesis in the domain of breast cancer surgery. We aimed at learning a structured model of operative reports for breast cancer using clustering methods applied to retrospective medical reports from CDW. We compared several clustering methods, and then selected the best for determining the set of items to include in breast surgery operative reports.

2. Methods

2.1. Data extraction and text processing

All data were extracted from the CDW of the Hôpital Européen Georges-Pompidou in Paris [7]. Patients were eligible if they had an operative report for breast cancer (ORBC) between 2010/01/01 and 2021/08/31 and if they had given their consent for research. The patients were identified as those having at least one ICD-10 code for breast cancer as primary diagnosis and at least one ORBC associated to a CCAM* code of breast surgery (*Classification Commune des Actes Médicaux). Patient consent has been provided by the Biological Resources Center and Tumor Bank Platform (BB-0033-00063), with the approval of the institution's ethic committee (IRB: #00011928).

Eligible ORBC were deduplicated, and administrative headings and footers were removed automatically. Each ORBC was then split into sentences, and carriage returns, punctuation, digits and French stopwords were removed. Any duplicated sentences were also removed to prevent the risk of clusters constituted of only one sentence.

2.2. Clustering methods

Each individual sentence was represented with TFIDF, Universal Language Encoder (USE), or Paraphrase model as proposed by the transformers library (PRP). These encoded sentences were then used as input for three clustering algorithms which have proven useful for text analysis in healthcare domain: k-means [4], DBSCAN [5], and Hierarchical Clustering [6]. This resulted in comparing 9 "couples" of token representation /clustering algorithm (3 token methods × 3 clustering algorithms). The best couple for determining the set of items to include in ORBC was selected using a two steps process. For each couple, we first determined the best number of clusters using grid search hyperparameters tuning. Then we calculated the relative size of each cluster compared to the total number of sentences. If one couple provided one noisy cluster

and/or one cluster containing more than 50% of the whole dataset of sentences, then the couple was excluded. Then, for each remaining couple, one author (AB) reviewed all the clusters. If a main theme could be identified in the considered cluster, then the cluster was considered as of good quality and was labelled (otherwise, it was not labelled). The coverage of each couple was then computed by calculating the ratio between the number of sentences included in labelled clusters and the total number of sentences in the dataset. The best couple was the one with the highest sentence coverage.

2.3. Determination of the set of items to include in ORBC

The labels of the clusters produced by the best couple were then reviewed and grouped by category, corresponding to the categories of items to include in ORBC.

3. Results

3.1. Comparison of clustering algorithms

1211 patients were eligible, resulting into a dataset of 1428 reports. From this dataset, 14105 unique sentences were used for clustering. The TFIDF/DBSCAN, PRP/DBSCAN, and USE/DBSCAN couples all produced one cluster containing more than 50% of the whole dataset of sentences, and as such were excluded per our exclusion criteria.

The TFIDF/K-means couple had the highest coverage (89%) and a high percentage of labelled clusters (91%) (Table 1).

Table 1. Comparison of clustering algorithms. Abbreviation used: HC: Hierarchical Clustering, KM: KMeans, PRP: Paraphrase algorithm, USE: Universal Sentence Encoder

	TFIDF/KM	**PRP/KM**	**USE/KM**	**TFIDF/HC**	**PRP/HC**	**USE/HC**
N of clusters	231	177	200	76	60	163
% of labelled clusters (n)	90 (208)	81 (144)	78 (156)	93 (71)	55 (33)	64 (105)
% of Sentence Coverage (n)	89 (12536)	77 (10833)	77 (10824)	80 (11289)	56 (7888)	63 (8844)

3.2. Set of items required for ORBC

From the clusters issued of TFIDF/K-means couple, we identified 7 categories of items to include in ORBC: (i) *administrative information*, containing 7 clusters (e.g., research consent); (ii) *medical history,* containing 19 clusters (e.g., disease discovery); (iii) *tumor characteristics*, containing 28 clusters (e.g., histology results); (iv) *tumor extension assessment*, containing 9 clusters (e.g., PET scan result); (v) *type of surgery*, containing 29 clusters (e.g., mastectomy); (vi) *medical staff involved in surgery*, containing 15 clusters (e.g., surgeon); (vii) *operative steps*, containing 101 clusters (e.g., hemostasis checking).

4. Discussion

We learnt a structured model of ORBC, using clustering algorithms applied on retrospective data extracted from the CDW of our hospital. TFIDF/k-means provided the

best coverage (91% of labelled clusters, and 89% of sentence coverage), and allowed us to learn a structured model of ORBC containing 7 main categories of items.

Our work has some limits. First, the selection of the best couple was not determined by comparing with a gold standard, but it is common in clustering and we tackled this issue by manually reviewing all the clusters obtaining results comparable with literature in which precision ranges from 52 to 97 % [3,8]. Second, the list of items retrieved depends on the quality of ORBC. We tried to limit this issue by considering a large retrospective set of reports written by various surgeons during a long period (10 years).

Surprisingly, we observed low performance of DBSCAN with the embedding models, in contrast with other studies demonstrating its efficacy in terms of sentence clustering [5] even outperforming K-Means [9]. This lack of performance could be explained in two ways: (i) the embeddings were not trained on medical datasets but on general wide datasets (USE) or paraphrase detection (PRP), (ii) DBSCAN automatically discard outliers into a "noise" cluster. An under-trained embedding could favor such a behavior by being too sparse. Re-training embeddings on a large dataset of medical reports extracted from the CDW could improve sentence coverage, as well as address French-specific syntax problems and medical domain usages. Nevertheless, results obtained with the TFIDF method, are highly satisfactory, resulting in an important coverage, and allow us to determine a set of items to include in operative breast reports.

We plan to implement the inferred structured model of operative report in the breast cancer surgery ward of our hospital for use in daily practice. This template will include the set of items identified in this study, as well as the related pre-established sentences found in each cluster. The provision of items and pre-established sentences, personalized to local practices, could increase adoption of operative template reports by surgeons, resulting in higher quality of operative reports and timesaving for surgeons.

References

[1] Tsopra R, Wyatt JC, Beirne P, Rodger K, Callister M, Ghosh D, et al. Level of accuracy of diagnoses recorded in discharge summaries: A cohort study in three respiratory wards. J Eval Clin Pract. 2019 Feb;25(1):36–43.

[2] Dean SM, Gilmore-Bykovskyi A, Buchanan J, Ehlenfeldt B, Kind AJ. Design and Hospital-Wide Implementation of a Standardized Discharge Summary in an Electronic Health Record. Jt Comm J Qual Patient Saf. 2016 Dec;42(12):555-AP11.

[3] Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics. 2019 Dec;132:103985.

[4] Naaz E, Sharma D, Sirisha D. Enhanced K-means Clustering Approach for Health Care Analysis Using Clinical Documents. 2016;8(1):5.

[5] Olago V, Bartels L, Dhokotera T, Bartels L, Bohlius J, Egger M, et al. The Use of Density-Based Spatial Clustering of Application With Noise (DBSCAN) for Record Linkage in An Observational HIV Cohort. IJPDS [Internet]. 2020 Dec 7 [cited 2022 Apr 8];5(5). Available from: https://ijpds.org/article/view/1422

[6] Koleck TA, Topaz M, Tatonetti NP, George M, Miaskowski C, Smaldone A, et al. Characterizing shared and distinct symptom clusters in common chronic conditions through natural language processing of nursing notes. Res Nurs Health. 2021 Dec;44(6):906–19.

[7] Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform. 2010;160(Pt 1):193–7.

[8] Wang Y, Tariq A, Khan F, Gichoya JW, Trivedi H, Banerjee I. Query bot for retrieving patients' clinical history: A COVID-19 use-case. Journal of Biomedical Informatics. 2021 Nov;123:103918.

[9] Mohammed SM, Jacksi K, Zeebaree SRM. Glove Word Embedding and DBSCAN algorithms for Semantic Document Clustering. In: 2020 International Conference on Advanced Science and Engineering (ICOASE) [Internet]. Duhok, Iraq: IEEE; 2020 [cited 2022 Jan 7]. p. 1–6. Available from: https://ieeexplore.ieee.org/document/9436540/