Clustering Italian medical texts: a case study on referrals

Clustering di testi medici italiani: un caso studio sulle ricette dematerializzate

Vittorio Torri, Michele Ercolanoni, Francesco Bortolan, Olivia Leoni and Francesca Ieva

Abstract In the medical domain, there is a large amount of valuable information that is stored in textual format. These unstructured data have long been ignored, due to the difficulties of introducing them in statistical models, but in the last years, the field of Natural Language Processing (NLP) has seen relevant improvements, with models capable of achieving relevant results in various tasks, including information extraction, classification and clustering. NLP models are typically language-specific and often domain-specific, but most of the work to date has been focused on the English language, especially in the medical domain. In this work, we propose a pipeline for clustering Italian medical texts, with a case study on clinical questions reported in referrals.

Key words: Natural Language Processing, Clustering, Administrative Databases, Medical Document

Vittorio Torri
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: vittorio.torri@polimi.it

Michele Ercolanoni
ARIA SpA (Azienda Regionale per L'Innovazione e gli Acquisti), Milan, Italy e-mail: michele.ercolanoni@ariaspa.it

Francesco Bortolan
UO Osservatorio Epidemiologico Regionale, Direzione Generale Welfare, Regione Lombardia, Milan, Italy e-mail: francesco_bortolan@regione.lombardia.it

Olivia Leoni
UO Osservatorio Epidemiologico Regionale, Direzione Generale Welfare, Regione Lombardia, Milan, Italy e-mail: olivia_leoni@regione.lombardia.it

Francesca Ieva
CHDS – Center for Health Data Science, Human Technopole, Milan, Italy
MOX - Modelling and Scientific Computing lab, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy e-mail: francesca.ieva@polimi.it

1 Introduction

In the last decade, Natural Language Processing (NLP) has seen significant advances and it has started to be applied also in the medical domain, where Electronic Health Records store nowadays a large number of textual documents, which often contain information that is not reported in the structured fields [5].

Medical documents present additional challenges with respect to those in the general domain, related to the specific lexicon, the high number of ambiguous abbreviations, the significant differences that exist in the way in which clinical documents are structured by different healthcare providers and the lack of annotated datasets.

Most of the work on NLP for the clinical domain is related to English documents, although in the last years, other languages are receiving increased attention. The extension of existing language models to different languages is not straightforward, due to the scarcity of (annotated) datasets and domain-specific resources [7]. In particular, there are only a few works which applied NLP to Italian medical documents, including both rule-based and machine-learning-based approaches [9, 6].

In this work, we present a pipeline for clustering Italian medical documents, applied to a specific case study: clinical questions that are reported in the referrals for specialized examinations.

In Italy, there is a standard format for referrals, containing various structured fields and two free-text fields: the *clinical question* and the *notes*. We focus on clinical questions since it is a field that must be compulsorily filled in with a description of the reason which determined the referral, i.e., the disease or the symptoms of the patient. For this purpose, we have been given access to a dataset of referrals filled-in in Lombardy Region in 2021, including their clinical questions.

Clinical questions are the only way to specify the reason for a referral since the Italian referral form does not provide any structured field for coding the disease or symptoms of the patients. Being able to automatically identify the disease code from the clinical questions would allow many subsequent analyses, both in the epidemiological field and on the appropriateness of prescriptions, two topics of paramount importance for the National Health System.

Referrals' clinical questions are not traditional clinical documents, in fact, they are administrative medical data. Administrative databases present the advantage of large coverage of the population, which is a relevant aspect for the above-mentioned types of analyses. Nevertheless, since they are meant neither for clinical nor for statistical use, they present some additional challenges with respect to traditional clinical documents: they often consist of one or two short sentences, missing punctuation and a proper syntactic structure, with multiple pieces of information not clearly separated (e.g., the proper clinical question, data on past examinations or past diseases, information on waiting times).

Due to the initial absence of an annotated dataset of clinical questions and since one of the most common problems in the medical domain is the absence of annotated datasets, we focused on the development of a clustering pipeline, which can be useful for different types of medical documents.

To the best of our knowledge, this is the first work on clustering of Italian referrals and more in general the first on clustering of Italian medical documents, with the exception of [1] which used proprietary software to cluster handovers of an Italian mental health institute.

The rest of the paper is organized as follows: in Section 2 we present the data and the clustering pipeline, in Section 3 we discuss preliminary results on the case study, together with limitations and possible future developments, while Section 4 contains conclusive remarks.

2 Materials and Methods

In this section we present the data for the case study and the clustering pipeline, discussing its main components.

2.1 Data

The dataset used in this work consists of 5000 clinical questions from referrals filled in by physicians in Lombardy Region in 2021, related to specialized oncological examinations. These data have been manually labelled with respect to the most common types of cancer, identifying 26 clusters, plus a cluster of all the remaining referrals. The distribution among the different types of cancer is severely unbalanced: breast cancer has 1287 referrals, the *other* group has 1144 referrals, while nine other types of cancers are in the order of hundreds and the remaining ones in the order of tens of referrals.

2.2 Clustering pipeline

Figure 1 depicts the pipeline with its main steps: pre-processing, text representation, dimensionality reduction and clustering. In this section, we analyze each of them.

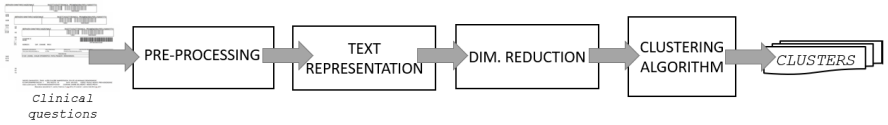

Fig. 1 Diagram of the clustering pipeline

2.2.1 Pre-processing

We considered the following steps: (i) lemmatization/stemmatization, (ii) abbreviations expansion, (iii) stop-words removal, (iv) typos correction, (v) low/high-frequency words removal. They are all common steps in NLP, with the exception of abbreviations expansion which is particularly relevant in the medical domain.

2.2.2 Text representation

In NLP there are two main types of representations that can be used: vocabulary-based representations and neural-network-based embeddings.

The most used vocabulary-based representation is the TF-IDF representation [2], where the entry for word w_j in document d_i of dataset D has the following value:

$$TF-IDF(w_j,d_i) = \frac{\text{\# occurrences of } w_j \text{ in } d_i}{\text{\# words in } d_i} \times log\left(\frac{|D|+1}{|d \in D : w_j \in d|}\right) \quad (1)$$

Vocabulary-based representations are sparse, high-dimensional and they do not take into account semantic similarity between words. Neural-network-based embeddings overcome this limitation since they produce fixed-length continuous vectors whose distance is related to the semantic similarity of the texts they represent. Here we focus on context-aware neural-network-based embeddings, i.e., those that are not fixed for a given word but depend also on the context in which it appears. BERT is a context-aware neural network model which has achieved state-of-the-art performances in many NLP tasks in the last years [4].

The original version of BERT has been developed for general-domain English documents, but many other versions of BERT have been subsequently developed to improve the performances on specific domains or different languages, using the same architecture but different training datasets. There are models for the medical domain [3], but they don't cover the Italian language, while there are models for the Italian language, but only for general-domain documents [8]. In our case study, we compare the results of Umberto, the most recent Italian version of BERT, with two other models that we have fine-tuned to cover the Italian medical domain:

1. **Umberto-medicina**: fine-tuned on documents extracted from the Medicine section of Italian Wikipedia (21.413 documents, 108 MB)
2. **Umberto-malattie**: fine-tuned on documents extracted from the Diseases section of Italian Wikipedia (4028 documents, 23 MB)

The datasets that we have used are much smaller than the original one (69GB), and for this reason, we adopted a transfer-learning approach, i.e. we fine-tuned Umberto instead of training a new BERT model from scratch.

2.2.3 Dimensionality Reduction

Vocabulary-based representations can easily reach hundreds of dimensions and even thousands. Neural-network-based embeddings have instead a fixed length, which is typically in the order of hundreds (768 for BERT). In both cases, there is a need to reduce their dimensionality before using them in a clustering algorithm. We applied PCA for this step, but other possibilities could be investigated in the future.

2.2.4 Clustering algorithm

The choice of the clustering algorithm depends on the type of representation, but also on the dataset. A relevant aspect, common to many clustering algorithms, is the distance metric to use. For vocabulary-based representations, cosine distance is a natural choice, while for neural-network-based embeddings both euclidean distance and cosine distance can be used.

3 Results

The results on the referrals dataset are measured in term of weighted F1-score, after having determined the best match between the clusters and the classes for which the data have been labelled. Table 1 shows a comparison between the TF-IDF representation and the different BERT-based embeddings. For TF-IDF we have used k-means as clustering algorithm, since it provided the best results, while for the BERT-based embeddings the best results have been achieved with HDBSCAN.

BERT-based language-models show large improvements with respect to the TF-IDF representation, and the fine-tuned versions of Umberto show a relevant improvement with respect to the standard Umberto. *Umberto-medicina* achieved the best results, probably because its fine-tuning dataset is larger than the one of *Umberto-malattie*.

Table 1 Results with different text representations

Text representation	W. Precision	W. Recall	W. F1
TF-IDF	0.167	0.223	0.184
Umberto	0.472	0.372	0.321
Umberto-medicina	**0.681**	**0.454**	**0.438**
Umberto-malattie	0.670	0.437	0.423

An analysis of the detailed results per type of cancer shows that that are significant differences between them. We have been able to identify clusters associated with 13 out of 26 cancer types. For most of the identified types of cancer, precision is high (> 0.90), while recall is low (< 0.50), with a few exceptions.

There are some types of cancers for which we identified multiple clusters that are associated with them. In particular, breast cancer, the largest class, has 5 clusters, while colon cancer has 3 clusters, in our best result. With TF-IDF the results are not only worse in terms of precision, recall and F1, but also in terms of number of clusters per cancer type, showing 9 different clusters for breast cancer. Some of these clusters referred to the same cancer type have a rationale for being separate, e.g., some of them contain indications of patients who have been under surgery, others are specifically referred to right/left breast, but this is not always the case. Further investigations are needed to understand how to avoid or merge these clusters.

4 Conclusions

In this work we have presented a pipeline for clustering Italian medical documents, addressing this problem for the first time. While many challenges remain open, relevant improvements have already been achieved with respect to baseline techniques, highlighting the benefits of BERT-based language models, even in this specific domain and without a large training dataset. Further investigations will be focused on the analysis of different dimensionality reduction techniques and on the problem of automatic merging different clusters related to the same topic.

References

1. Accordini, Monica, et al. Stories of change: The text analysis of handovers in an Italian psychiatric residential care home. Journal of Psychiatric and Mental Health Nursing 24.4, 232-242 (2017)
2. Aggarwal, Charu C. Mining text data. Data mining. Springer, Cham, (2015)
3. Alsentzer, Emily, et al. Publicly Available Clinical BERT Embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop (2019)
4. Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 4171-4186 (2019)
5. Iroju, Olaronke G., and Janet O. Olaleke. A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science 8, 44-50 (2015)
6. Lanera, Corrado, et al. Use of machine learning techniques for case-detection of varicella zoster using routinely collected textual ambulatory records: Pilot observational study. JMIR Medical Informatics 8.5, e14330 (2020)
7. Névéol, Aurélie, et al. Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics 9.1, 1-13 (2018)
8. Tamburini, Fabio. How "BERTology" Changed the State-of-the-Art also for Italian NLP. CLiC-it (2020)
9. Viani, Natalia, et al. Information extraction from Italian medical reports: An ontology-driven approach. International journal of medical informatics 111, 140-148 (2018)