Injection of Automatically Selected DBpedia Subjects in Electronic Medical Records to boost Hospitalization Prediction

Raphaël Gazzotti, Catherine Faron Zucker, Fabien Gandon, Virginie Lacroix-Hugues, David Darmon

Injection of Automatically Selected DBpedia Subjects in Electronic Medical Records to boost Hospitalization Prediction

Raphaël Gazzotti
Université Côte d'Azur, Inria, CNRS,
I3S, Sophia-Antipolis, France
SynchroNext, Nice, France
raphael.gazzotti@unice.fr

Catherine Faron-Zucker
Université Côte d'Azur, Inria, CNRS,
I3S, Sophia-Antipolis, France
catherine.faron@unice.fr

Fabien Gandon
Inria, Université Côte d'Azur, CNRS,
I3S, Sophia-Antipolis, France
gandon.fabien@inria.fr

Virginie Lacroix-Hugues
Université Côte d'Azur, RETINES,
Département d'Enseignement et de
Recherche en Médecine Générale,
Faculté de médecine, Nice, France
vhugues@outlook.fr

David Darmon
Université Côte d'Azur, RETINES,
Département d'Enseignement et de
Recherche en Médecine Générale,
Faculté de médecine, Nice, France
david.darmon@unice.fr

ABSTRACT

Although there are many medical standard vocabularies available, it remains challenging to properly identify domain concepts in electronic medical records. Variations in the annotations of these texts in terms of coverage and abstraction may be due to the chosen annotation methods and the knowledge graphs, and may lead to very different performances in the automated processing of these annotations. We propose a semi-supervised approach based on DBpedia to extract medical subjects from EMRs and evaluate the impact of augmenting the features used to represent EMRs with these subjects in the task of predicting hospitalization. We compare the impact of subjects selected by experts vs. by machine learning methods through feature selection. Our approach was experimented on data from the database PRIMEGE PACA that contains more than 600,000 consultations carried out by 17 general practitioners (GPs).

CCS CONCEPTS

• **Applied computing → Health informatics**; • **Theory of computation → Semantics and reasoning**; • **Computing methodologies → Information extraction**; Feature selection;

KEYWORDS

Information extraction, Predictive model, Electronic medical record, Knowledge graph.

ACM Reference Format:
Raphaël Gazzotti, Catherine Faron-Zucker, Fabien Gandon, Virginie Lacroix-Hugues, and David Darmon. 2020. Injection of Automatically Selected DBpedia Subjects in Electronic Medical Records to boost Hospitalization Prediction. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3341105.3373932

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373932

1 INTRODUCTION

Electronic medical records (EMRs) contain vital information about a patient's state of health, and their analysis should enable preventing pathologies that may affect a patient in the future. Their exploitation through automated approaches makes it possible to discover patterns that, once addressed, are likely to improve the living conditions of the population. However, linguistic variability and tacit knowledge hinder automated processing, as they can lead to erroneous conclusions. In this paper we extract entities that help predict the hospitalization of patients from their electronic medical records and linked DBpedia[1] entities. DBpedia employs Semantic Web standards and structures Wikimedia project data with the Resource Description Framework (RDF). However, given the amount of general information available on DBpedia, it is challenging to filter knowledge specific to the healthcare domain. This is especially the case when it comes to identify concept relevant to the prediction of hospitalized patients. To answer this problem, we estimate the relevance of concepts and select the most promising ones to construct the vector representation of EMRs used to predict hospitalization.

As a field of experimentation, we used a dataset extracted from the PRIMEGE PACA relational database [15] which contains more than 600,000 consultations in French by 17 general practitioners (Table 1). In this database, text descriptions written by general practitioners are available with international classification codes of prescribed drugs, pathologies and reasons for consultations, as well as the numerical values of the different medical examination results obtained by a patient.

In that context, our main research question is: How to extract knowledge relevant for the prediction of the occurrence of an event? In our case study, we extract subjects related to hospitalization using knowledge from DBpedia and EMRs. In this paper, we focus on the following sub-questions:

- How to filter relevant domain knowledge from a general knowledge source?
- How to deal with subjectivity in the annotation process?

[1] DBpedia is a crowd-sourced extraction of structured data from Wikimedia projects http://dbpedia.org

Table 1: Data collected in the PRIMEGE PACA database.

Category	Data collected
GPs	Sex, birth year, city, postcode
Patients	Sex, birth year, city, postcode Socio-professional category, occupation Number of children, family status Long term condition (Y/N) Personal history Family history Risk factors Allergies
Consultations	Date Reasons of consultation Symptoms related by the patient and medical observation Further investigations Diagnoses Drugs prescribed (dose, number of boxes, reasons of the prescription) Paramedical prescriptions (biology/imaging) Medical procedures

- Is automatic extraction and selection of knowledge efficient in that context?

To answer these questions, we survey the related work (section 2) and position our contribution. We then introduce the proposed method for knowledge extraction from texts and specify the filters used to retrieve medical knowledge (section 3). Subsequently, we present the experimental protocol to compare the impact of knowledge selected by experts and automatic selection, and we discuss the results obtained (section 4). Finally, we conclude and provide our perspectives for this study (section 5).

2 RELATED WORK

In [6], to address data insufficiency and interpretation of deep learning models for the prediction of rarely observed diseases, the authors established a neural network with graph-based attention model that exploits ancestors extracted from the OWL-SKOS representations of ICD Disease, Clinical Classifications Software (CCS) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). In order to exploit the hierarchical resources of these knowledge graphs in their attention mechanism, the graphs are transformed using GloVe embeddings [19]. The results show that the proposed model outperforms a standard recurrent neural network when identifying pathologies that are rarely observed in the training data, at the same time also generalising better when only few training instances are available.

In [20], to improve accuracy in the recognition of daily living activities, the authors extract knowledge from the dataset of [17] and structure it with a knowledge graph developed for this purpose. Then, they automatically deduce new class expressions, with the objective of extracting their attributes to recognize activities of daily living using machine learning algorithms. The authors highlight better accuracy and results than with traditional approaches, regardless of the machine learning algorithm on which this task has been addressed (up to 1.9% on average). Although they exploit solely the knowledge graph developed specifically for the purpose of discovering new rules, without trying to exploit other knowledge sources where a mapping could have been done. Their study shows the value of structured knowledge in classification tasks.

The SIFR Bioportal project [21] provides a web service based on the NCBO BioPortal [23] to annotate clinical texts in French with biomedical knowledge graphs. This service is able to handle clinical notes involving negations, experiencers (the patient or members of his family) and temporal aspects in the context of the entity references. However, the adopted approach involves domain specific knowledge graphs, while general resources like EMRs require general repositories such as, for instance, DBpedia.

In [10], the authors show that combining bag-of-words (BOW), biomedical entities and UMLS (the Unified Medical Language System[2]) improve classification results in several tasks such as information retrieval, information extraction and text summarization regardless of the classifier. We intend here to study the same kind of impact but from a more general repository like DBpedia and on a domain-specific prediction task: we also propose a method to select relevant domain knowledge in order to boost hospitalization prediction.

In [11], we studied the contributions of different knowledge graphs (ATC, ICPC-2, NDF-RT, Wikidata and DBpedia) for hospitalization prediction. Compared to [11], this paper explores in more depth the impact of knowledge enrichment using DBpedia while relying on the same prediction method. Our goal is to provide a method to solve the problem of retrieving relevant knowledge in the medical domain from general knowledge source. The intuition behind the use of DBpedia is that general knowledge is only available on general repositories, and the way knowledge is structured differs from specialized referentials. To achieve this purpose, we propose a method that relies on semi-supervised learning to extract subject candidates. Selecting concepts relevant for a specific domain problem is both an expert and subjective task [12] for which an automated solution could help develop new applications.

3 KNOWLEDGE EXTRACTION AND REPRESENTATION OF EMR

3.1 Extraction of candidate subjects from DBpedia to predict hospitalization

The first step of our approach consists in recognizing named entities from the medical domain part of DBpedia within French texts contained in EMRs. This is performed by an instance of the semantic annotator DBpedia Spotlight [8] that was deployed locally and pretrained with a French model.[3] To ensure that the retrieved entities belong to the medical domain, we enforce two constraints on the resources identified by DBpedia Spotlight. The first constraint requires that the identified resources belong to the medical domain of the French chapter of DBpedia. The second one does the same

[2] UMLS is a metathesaurus developed at the US National Library of Medicine http://www.nlm.nih.gov/pubs/factsheets/umls.html

[3] https://sourceforge.net/projects/dbpedia-spotlight/files/2016-10/fr/

Listing 1: SPARQL query to extract subjects related to the medical domain from DBpedia.

```
1  PREFIX dbo: <http://dbpedia.org/ontology/>
2  PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
3  PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
4  PREFIX dcterms: <http://purl.org/dc/terms/>
5  PREFIX yago: <http://dbpedia.org/class/yago/>
6  PREFIX cat: <http://fr.dbpedia.org/resource/Catégorie:>
7
8  SELECT ?skos_subject WHERE {
9    SERVICE <http://fr.dbpedia.org/sparql> {
10     # Constraint on the medical domain
11     VALUES ?concept_constraint {
12      cat:Maladie             # disease
13      cat:Santé               # health
14      cat:Génétique_médicale  # medical genetics
15      cat:Médecine            # medicine
16      cat:Urgence             # urgency
17      cat:Traitement          # treatment
18      cat:Anatomie            # anatomy
19      cat:Addiction           # addiction
20      cat:Bactérie            # bacteria
21     }
22     <link_dbpedia_spotlight> dbpedia-owl:wikiPageRedirects{0,1} ?page.
23     ?page dcterms:subject ?page_subject.
24     ?page_subject skos:broader{0,10} ?concept_constraint.
25     ?page_subject skos:prefLabel ?skos_subject.
26     ?page owl:sameAs ?page_en.
27     # Filter used to select the corresponding resource in the English Chapter of
           DBpedia
28     FILTER(STRSTARTS(STR(?page_en), "http://dbpedia.org/resource/"
           ))
29   }
30
31   SERVICE <http://dbpedia.org/sparql> {
32     VALUES ?type_constraint {
33      dbo:Disease
34      dbo:Bacteria
35      yago:WikicatViruses
36      yago:WikicatRetroviruses
37      yago:WikicatSurgicalProcedures
38      yago:WikicatSurgicalRemovalProcedures
39     }
40     ?page_en a ?type_constraint
41   }
42 }
```

with the English chapter in order to filter and select health domain-related subjects and to overcome the defects of the French version in which property `rdf:type` is poorly used. This involves calling two SERVICE clauses in a SPARQL query,[4] each one implementing a constraint according to the structure of the French and English chapter it remotely queries. The workflow is represented in Figure 1 and the query in Listing 1.

From the URIs of the identified resources, the first part of the query (lines 9-29) accesses the French chapter of DBpedia to check that the value of their property `dcterms:subject`[5] belongs to one of the hierarchies of SKOS concepts (`skos:broader`, `skos:narrower`) having for roots the French terms for disease, health, medical genetics, medicine, urgency, treatment, anatomy, addiction and bacteria.

The second part of the query (lines 31-41) checks that the identified resources from the French DBpedia have for its English equivalent (`owl:sameAs`) at least one of the following types

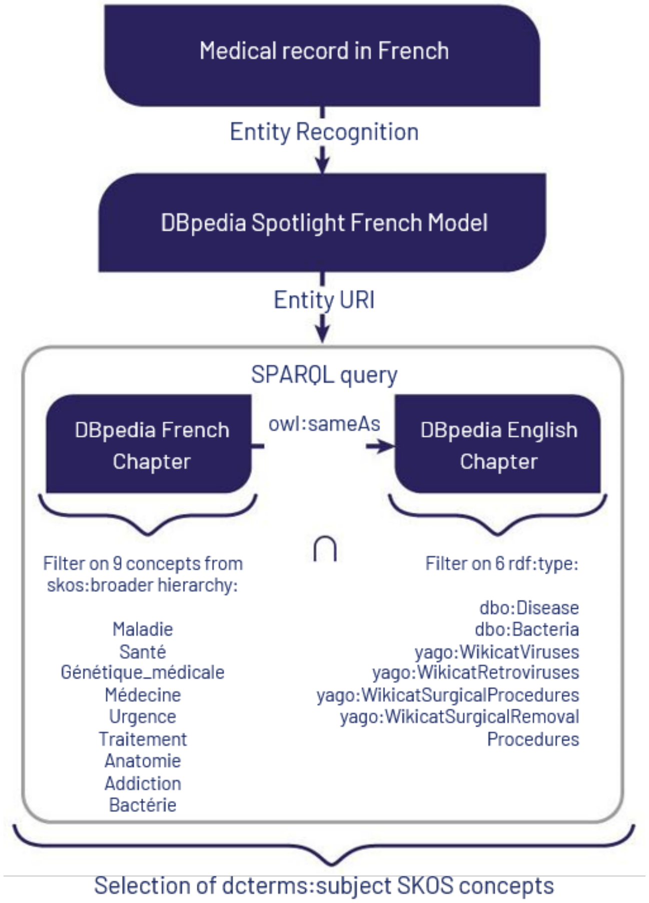

Figure 1: Workflow used to extract candidate subjects from EMR.

(`rdf:type`)[6]: dbo:Disease, dbo:Bacteria, yago:WikicatViruses, yago:WikicatRetroviruses, yago:WikicatSurgicalProcedures, yago:WikicatSurgicalRemovalProcedures.

We do not consider some other types like `dbo:Drug`, `dbo:ChemicalCoumpound`, `dbo:ChemicalSubstance`, `dbo:Protein`, or `yago:WikicatMedicalTreatments`, as they generate answers related to chemical compounds: the retrieved resources can thus range from drugs to plants, to fruits. We do not consider either types referring to other living beings like `umbel-rc:BiologicalLivingObject` or `dbo:Species` which are too general to return relevant results. We do not consider either many biomedical types in the yago namespace which URI ends by an integer (e.g., http://dbpedia.org/class/yago/Retrovirus101336282), which are too numerous and too close from each other. The type `dbo:AnatomicalStructure` is also non-relevant with this second constraint since it retrieves subjects related to different anatomical parts which are not human specific. The list of labels of concepts thus extracted allows to construct a vector representation of EMRs used to identify hospitalized patients.

[4] https://www.w3.org/TR/sparql11-query/
[5] Namespace: http://purl.org/dc/terms/
[6] Namespaces: http://dbpedia.org/ontology/, http://dbpedia.org/class/yago/

In order to improve DBpedia Spotlight's detection capabilities, words or abbreviated expressions within medical reports are added to text fields using a symbolic approach, with rules and dictionaries. For instance the abbreviation "ic" which means "heart failure" is not recognized by DBpedia Spotlight, but is correctly identified by our rule-based approach. This method to retrieve classification labels was applied on all the textual fields of our dataset.

3.2 Injection of concepts in the vector representation of EMRs

A domain specific corpus like PRIMEGE contains very specialized jargon, and which have a meaning adapted to its context. This is why we chose to use a bag-of-words (BOW) representation to avoid out of vocabulary issues. It allows us to generate our own textual representation of EMRs since it does not require a large amount of data. This model also enables to identify the contribution of each term to distinguish patients to hospitalize or not, thus answering algorithm explanation issues. Additionally, the integration of heterogeneous data is facilitated since it is sufficient to concatenate other attributes to this model without removing the meaning of the terms previously represented in this way.

Moreover, loss of information is intrinsic to more advanced data representation models. We have opted for a BOW in order to remain able to provide general practitioners with the closest information available in their files.

In order to mimic the structure of the PRIMEGE database and to prevent wrong conclusions, we have introduced provenance prefixes during the creation of the bag-of-words to trace the contribution of the different fields. This allows to distinguish some textual data from each other in the vector representation of EMRs, e.g., a patient's personal history and his family history.

Subjects label from DBpedia are considered as a token in a textual message. When an entity is identified in a patient's medical record, the label of his corresponding subject is added to a concept vector. This attribute will have as value the number of occurrences of this subject within the patient's health record (e.g., the subjects 'Organ failure' and 'Medical emergencies' -among other concepts- are identified for 'pancréatite aiguë', acute pancreatitis, and the value for these attributes in our concept vector will be equal to 1).

Let $V^i = \{w_1^i, w_2^i, ..., w_n^i\}$ be the bag-of-words obtained from the textual data in the EMR of the i^{th} patient. Let $C^i = \{c_1^i, c_2^i, ..., c_m^i\}$ be the bag of concepts for the i^{th} patient resulting from the extraction of labels of concepts belonging to DBpedia after analysis of his consultations from semi-structured data such as text fields listing drugs, pathologies, and unstructured data from free texts such as observations. The vector representation of the i^{th} patient is the sum of V^i and C^i or a sub-vector of it, as detailed in the next section. Figure 2 represents the general workflow used to generate vector representations.

3.3 Alternative vector representations: manual vs. automatic selection of relevant subjects

To decide on the optimal vector representation of a patient's EMR, we considered further filtering the list of the labels of concepts extracted from DBpedia, depending on their relevancy for the targeted prediction task. We first submitted the list of the 285 extracted labels of concepts to human medical experts who were asked to assess their relevance for studying patients' hospitalization risks from their EMRs. Alternatively, we considered automatically selecting the concepts relevant for studying hospitalization by using a feature selection algorithm applied on a training set of vector representations of patients in the C^i form.

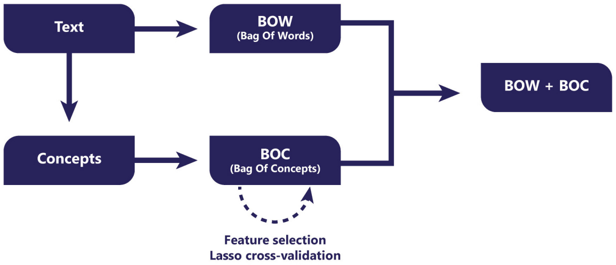

Figure 2: Workflow used to generate vector representations integrating ontological knowledge alongside with textual information.

As a result we generated the following alternative vector representations that should be compared when used to predict hospitalization.

- *baseline*: represents our basis of comparison where no enrichment with DBpedia concepts is made on EMR data, i.e., only text data in the form of bag-of-words: V^i
- α: refers to an enrichment of V^i with the labels of concepts automatically extracted from the DBpedia knowledge base: $V^i + C^i$.
- β: refers to an enrichment of V^i with a subset of the labels of concepts in C^i acknowledged as relevant by at least one expert human annotator.
- γ: refers to an enrichment of V^i with a subset of the labels of concepts in C^i acknowledged as relevant by all the experts human annotators.
- ϵ: refers to an enrichment of V^i with a subset of the labels of concepts in C^i output by the automatic feature selection algorithm. We chose the Lasso algorithm [22] and we executed it *within* the internal loop of the nested cross-validation in the global machine learning algorithm chosen to predict hospitalization. For the Lasso algorithm, we chose the default parameters (and the number of folds used for cross-validating in that context, fixed at $F = 3$).

4 EXPERIMENTS AND RESULTS

4.1 Dataset and protocol

The extraction of labels of concepts described in Section 3.1 was performed on a sample of 1446 patients, DS_B, a balanced dataset. This dataset contains data on 714 patients hospitalized and 732 patients not hospitalized. Then, we introduce these concepts in the vector representation of EMRs with the same dataset and classify them to predict the future hospitalization of patients or not.

To construct V^i we consider the following EMR fields: sex, birth year, long term condition, risk factors, allergies, reasons of consultation with their associated codes, medical observations, diagnoses with their associated codes, care procedures, drugs prescribed with their associated codes and reasons of the prescription, patient's history, the family history, past problems and symptoms of the patient. Most of the concepts in C^i are extracted from the field "reasons of consultation" that is very short and the field "medical observations" which length generally goes from 50 to 300 characters. By default, C^i does not use the following fields: patient's history, the family history, past problems and symptoms of the patient.

An alternative vector representation of EMRs, ζ, uses for C^i the following additional fields: patient's history, the family history, past problems and symptoms of the patient are processed. Note that the symptom field as the observation field is used by physicians for various purposes. ζ is constructed similarly to ϵ when considering these additional data.

Since we evaluate our vector representation with non-sequential machine learning algorithms, we aggregated all patients' consultations to overcome the temporal dimension specific to EMRs. All consultations occurring before hospitalization are aggregated into a vector representation of the patients' medical file. For patients who have not been hospitalized, all their consultations are aggregated. Thus, the text fields contained in patients' records are transformed into vectors.

We evaluated the vector representations by nested cross-validation [4], with an outer loop with a $K = 10$, and an inner loop with $L = 3$. The exploration of hyperparameters was performed by random search [2] over 150 iterations.

The different experiments were conducted on an HP EliteBook 840 G2, 2.6 GHz, 16 GB RAM with a virtual environment under Python 3.6.3. The creation of vector representations was done on the HP EliteBook and on this same machine were deployed DBpedia Spotlight and Corese Semantic Web Factory [7],[7] a software platform for the Semantic Web that implements RDF, RDFS, SPARQL 1.1 Query & Update, and OWL RL.

4.2 Inter-rater reliability of subject annotation

Two general practitioners and one biologist have independently annotated the 285 subjects extracted from DBpedia. The annotations were transformed in vectors with a size of 285. Then, we compared with the Krippendorff's α metric vectors resulting from human annotation. We compare up to 3 vectors with Krippendorff's α metric, i. e. with the biologist and physicians, and up to 2 vectors to compare only the physicians' annotations. The correlation metric was used to compare different pairs of vectors resulting either from human or machine annotation. The Figure 3 shows the workflow used to assess inter-rater reliability.

Annotations have been evaluated towards the Krippendorff's α metric [14] and obtained a score of 0.51, the annotation score between the two general practitioners is of 0.27.

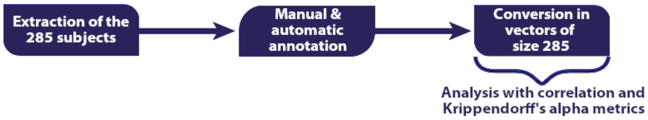
Extraction of the 285 subjects
Manual & automatic annotation
Conversion in vectors of size 285
Analysis with correlation and Krippendorff's alpha metrics

Figure 3: Workflow used to compute inter-rater reliability for both human and machine annotations.

Even by excluding some subjects involving a terminological conflict in their naming, since if someone annotates the beginning of a label of concept as relevant towards the hospitalization of a patient (the opposite is also true) all the labels of concepts starting with the same expression will be annotated in the same way. In doing so, the three annotators obtained a score of 0.66, and 0.52 for the inter-rater reliability between the two general practitioners. The subjects excluded started by 'Biology', 'Screening and diagnosis', 'Physiopathology', 'Psychopathology', 'Clinical sign', 'Symptom' and 'Syndrome' which brings us back to a new total of 243 concepts.

On average, 198 subjects were annotated by experts as relevant to the study of patients' hospitalization risks, respectively 217 and 181 for the general practitioners and 196 for the biologist among the 285 subjects proposed with the extraction based on the SPARQL query displayed in Section 3.1.

As discussed by [1], a score within this range of values is insufficient to drawn conclusions and it shows the difficulty of this task, both because identifying entities involved in patient hospitalization is subject to interpretation and because it is complex to find consensus in this task that could be seen at first sight as simplistic by an expert in the field.

The union of labels of concepts identified with the ζ approach counts 51 different subjects (63 if the provenance prefix is considered as a different subjects) and the intersection of labels of concepts identified with ζ counts 14 different subjects (19 if the provenance prefix is considered as a different subject). Table 2 displays correlation metric values between experts and machine annotators (its value ranges from 0 to 2, meaning that 0 is a perfect correlation, 1 no correlation and 2 perfect negative correlation). This metric was computed by comparing among the 285 subjects, if they are deemed relevant, irrelevant or not annotated (in the case of human annotation) to study the patient's hospitalization risks from their EMRs, thus vectors are compared in pairs in this table.

Table 2 shows up a wide variation between human annotators and machine annotators (maximum of 1.1399 between A_1 and M_4), whereas between annotators of a specific group this margin is not significant (maximum of 0.6814 for humans and maximum of 0.4185 for machines). The union of subjects U_1 retrieved by machine annotators is really similar to M_5, since they have a correlation score of 0.12.

4.3 Selected machine learning algorithms

We performed the hospitalization prediction task with different state of the art algorithms available in the Scikit-Learn library [18]. The optimized hyperparameters determined by nested cross-validation are as follows:

- *SVC*, C-Support Vector Classifier, which implementation is based on the libsvm implementation [5]: The regularization coefficient C, the kernel used by the algorithm and the gamma coefficient of the kernel.
- *RF*, Random Forest classifier [3]: The number of trees in the forest, the maximum depth in the tree, the minimum number

[7] http://corese.inria.fr

Table 2: Correlation metric ($1 - \frac{(u-\bar{u}).(v-\bar{v})}{\|u-\bar{u}\|_2 \|v-\bar{v}\|_2}$, with $\bar{u}$, the mean of elements of u, and respectively $\bar{v}$, the mean of elements of v) computed on the 285 subjects. A_1 to A_3 refers to human annotators and M_1 to M_{10} refers to machine learning through feature selection annotation on the ζ approach (considering the 10 K-Fold). U_1 is the union of subjects from the sets M_1 to M_{10}.

	A_1	A_2	A_3	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	U_1
A_1	\	0.6814	0.4180	1.1085	1.0688	1.1138	1.1399	1.0692	1.1166	1.1085	1.0688	1.1257	1.1363	1.1405
A_2	0.6814	\	0.2895	1.0618	1.1066	1.0072	1.0745	1.0534	1.1127	1.0618	1.0611	1.0904	1.0749	1.0737
A_3	0.4180	0.2895	\	1.0232	1.0807	1.0242	1.0721	1.0616	1.0708	1.0232	1.0320	1.0708	1.0520	1.0933
M_1	1.1085	1.0618	1.0232	\	0.2105	0.2635	0.2249	0.3410	0.3389	0.2116	0.2105	0.2031	0.2760	0.3293
M_2	1.0688	1.1066	1.0807	0.2105	\	0.2319	0.1605	0.1597	0.2037	0.1714	0.0724	0.2358	0.3019	0.2605
M_3	1.1138	1.0072	1.0241	0.2635	0.2319	\	0.1408	0.2700	0.2865	0.2249	0.1605	0.3346	0.2710	0.2472
M_4	1.1399	1.0745	1.0721	0.2249	0.1605	0.1408	\	0.2700	0.2527	0.1863	0.1248	0.2495	0.2710	0.2472
M_5	1.0692	1.0534	1.0616	0.3410	0.1597	0.2700	0.2700	\	0.2508	0.2379	0.1597	0.3595	0.4167	0.1200
M_6	1.1166	1.1127	1.0708	0.3389	0.2037	0.2865	0.2527	0.2508	\	0.2275	0.2037	0.3690	0.3495	0.2080
M_7	1.1085	1.0618	1.0232	0.2116	0.1714	0.2249	0.1863	0.2379	0.2275	\	0.1322	0.1565	0.3238	0.3293
M_8	1.0688	1.0611	1.0320	0.2105	0.0724	0.1605	0.1248	0.1597	0.2037	0.1322	\	0.2358	0.3019	0.2605
M_9	1.1257	1.0904	1.0708	0.2031	0.2358	0.3346	0.2495	0.3595	0.3690	0.1565	0.2358	\	0.2888	0.4030
M_{10}	1.1363	1.0749	1.0520	0.2760	0.3019	0.2710	0.2710	0.4167	0.3495	0.3238	0.3019	0.2888	\	0.4185
U_1	1.1405	1.0737	1.0933	0.3293	0.2605	0.2472	0.2472	0.1200	0.2080	0.3293	0.2605	0.4030	0.4185	\

of samples required to split an internal node, the minimum number of samples required to be at a leaf node and the maximum number of leaf nodes.

- *Log*, Logistic Regression classifier [16]: The regularization coefficient C and the penalty used by the algorithm.

One of the motivations for using these algorithms is because logistic regression and random forest are widely used in order to predict risk factors in EMR [13]. These machine learning algorithms are able to provide a native interpretation of their decisions. The reasons leading to a patient's hospitalization are thus reported to the physician, as well as the factors on which the physician can intervene to prevent this event from occurring. Moreover, the limited size of our dataset excluded neural networks approaches.

4.4 Results

We used the $F_{tp,fp}$ metric [9] to evaluate the performance of machine learning algorithms. Let TN be the number of negative instances correctly classified (True Negative), FP the number of negative instances incorrectly classified (False Positive), FN the number of positive instances incorrectly classified (False Negative) and TP the number of positive instances correctly classified (True Positive). K represents the number of loops used to cross-validate (in our context this number is fixed at 10) and the notation $_f$ is used to distinguish a fold related metric like the amount of true positives to the sum of true positives across all folds.

$$TP_f = \sum_{i=1}^{K} TP^{(i)} \qquad FP_f = \sum_{i=1}^{K} FP^{(i)} \qquad FN_f = \sum_{i=1}^{K} FN^{(i)}$$

$$F_{tp,fp} = \frac{2.TP_f}{2.TP_f + FP_f + FN_f}$$

The comparison of the different features sets is presented in Table 3. Results with only bag of concepts were not included (no

Table 3: $F_{tp,fp}$ for the different vector sets considered on the balanced dataset DS_B.

Features set	*SVC*	*RF*	*Log*	Average
baseline	0.8270	**0.8533**	0.8491	0.8431
α	0.8214	0.8492	0.8388	0.8365
β	0.8262	0.8521	0.8432	0.8405
γ	0.8270	0.8467	0.8445	0.8394
ϵ	0.8363	0.8547	**0.8642**	0.8517
ζ	0.8384	0.8541	**0.8689**	0.8538

Table 4: Confusion matrix of the random forest algorithm (on the left) and the logistic regression (on the right) on the *baseline* ('H' stands for Hospitalized and 'Not H' for 'Not Hospitalized').

	H	Not H
Predicted as 'H'	599	91
Predicted as 'Not H'	115	641

	H	Not H
Predicted as 'H'	588	83
Predicted as 'Not H'	126	649

Table 5: Confusion matrix of ζ (on the left) and the union of subjects under ζ conditions (on the right) approaches under the logistic regression algorithm ('H' stands for Hospitalized and 'Not H' for 'Not Hospitalized').

	H	Not H
Predicted as 'H'	600	67
Predicted as 'Not H'	114	665

	H	Not H
Predicted as 'H'	603	67
Predicted as 'Not H'	111	665

feature from the baseline), since with the 285 subjects, C^i, and the logistic regression algorithm we obtained a $F_{tp,fp}$ of 0.6778.

4.5 Generalization of concepts vector

Following the list of labels of concepts extracted for each fold with the ζ approach, we evaluate the effect of a global vector of concepts since with our experimentation setup the selected features can be different from one fold to another, i.e., a same vector of concepts across all folds. Thus, we generate different stable vector of concepts based on the number of intersections of subjects and union of subjects with the hyperparameters identified with the ζ approaches.

The intersection of all the subjects gets a score of 0.8662 and the union of all subjects encountered for each fold obtains a score of 0.8714 which is better than the baseline (by more than 2%) and even better than the ζ approach. For the generalization of concepts vector, the main gain is shown on the right of the Table 5 with the increase of true positives and therefore reduction of false negatives.

4.6 Discussion

The SPARQL query in Listing 1 allowed to extract a list of medical subjects from DBpedia considered relatively relevant to the issue of hospitalization since approximately 198 subjects out of 285 were annotated in this way by experts.

The best performing approach, ζ, selected a much smaller number of subjects with a feature selection process, this implies that the selected subjects are more precise in order to distinguish hospitalized patients from other ones (Tables 4 and 5) by improving both the detection of true positives and true negatives. The union of subjects also improves the number of true positives in comparison to the ζ approach. That means that a step involving a feature selection algorithm allows to retrieve the most relevant labels of concepts in a context where the training dataset is small and may help with annotation procedures. Although this requires a more specific selection, comparing the results obtained with ϵ and ζ approach shows that subjects not directly related to the patient's own case helps to predict his hospitalization.

Among the 51 labels of concepts selected with the union of subjects, more generic knowledge was selected like 'Terme médical' (respectively 'Medical Terminology'), one possibility could be that the general practitioner uses a technical terminology in a situation involving a complex medical case. Numerous concepts related to patient's mental state (like 'Antidépresseur', 'Dépression (psychiatrie)', 'Psychopathologie', 'Sémiologie psychiatrique', 'Trouble de l'humeur') appear to be a cause of hospitalization. Different concepts related to the allergy ('Allergologie', 'Maladie pulmonaire d'origine allergique') and infectious diseases ('Infection ORL', 'Infection urinaire', 'Infection virale', 'Virologie médicale') were selected. Concepts related to the cardiovascular system are widely represented within this set ('Dépistage et diagnostic du système cardio-vasculaire', 'Maladie cardio-vasculaire', 'Physiologie du système cardio-vasculaire', 'Signe clinique du système cardio-vasculaire', 'Trouble du rythme cardiaque'). The only concept retrieved in the family history of the patient, at the exception of 'Medical Terminology', is 'Diabète' (respectively 'Diabetes'). Among the labels of concepts selected by machine learning through feature selection, rare concepts considered irrelevant at first sight toward the problem of hospitalization such as 'Medical Terminology' could find an explanation. Also, a feature selection step helps to improve the prediction of hospitalization by adding knowledge indirectly related to the patient's condition, such as family history (approach ζ).

Although the number of subjects considered as relevant by experts is quite high, their integration into a vector representation reduced the performance obtained in comparison to the baseline, one of the possibilities for this result is the limited size of our annotated corpus. One of the weaknesses of this approach is that a knowledge base like DBpedia may be incomplete (incompleteness of properties `dcterms:subjects`, `owl:sameAs` and `rdf:type`), which would justify in order to obtain better results to proceed to the content curation of such knowledge base.

The incompleteness of medical records implies a huge variety between patient and from one consultation to another for the same patient according to the level of information provided by the general practitioner. Also, joint medical care by a fellow specialist with sometimes little information about these cares is another negative factor. Moreover, the patient may not have been detected as being particularly at risk or may not be very observant and does not come a lot to consultations, this shows the interest of being able to work on patient trajectories and to set up a health data warehouse combining several sources.

Reports of the consultations contain abbreviations of experts and thus it would lead to significant improvements in the knowledge extraction task to be able to distinguish abbreviations and their meanings in a given medical context. We plan to detect negation and experiencer in future work since a pathology affecting a patient's relationship or the negation of a pathology does not carry the same meaning when it comes to predict a patient's hospitalization.

5 CONCLUSION

In this paper, we have presented a method to extract from the DBpedia knowledge base, subjects related to the medical domain. Then, we evaluated their performance with different machine learning algorithms to predict hospitalization when they are injected in the vector representation of EMRs. Deciding the relevancy of given subjects for a specific prediction task appeared to be quite difficult and subjective for human experts, with a high variability in their annotations. To overcome this problem, we integrated an automatic step allowing annotators to confirm their thoughts. We generated different vector representations coupling concepts vectors and bag-of-words and then evaluated their performance for prediction with different machine learning algorithms and computed inter-rater reliability metrics for different sets of concepts whether selected by the human or the machine. Our contributions are in the automatic extraction of DBpedia subjects and injection of the latter into EMRs representation, the coupling with a feature selection method to select relevant resources towards hospitalization risks, the selection and evaluation of subjects by both human and machine annotators.

As future work, we plan to train our own model of DBpedia Spotlight in order to further avoid noise with named entities from other domains. We also intend to investigate different depth levels of subjects, since so far, we only integrated the knowledge on the direct subject, and to deal with the recognition of complex expressions, experiencer and entity negation.

ACKNOWLEDGMENTS

This work is partly funded by the French government labelled PIA program under its IDEX UCAJEDI project (ANR-15-IDEX-0001).

REFERENCES

[1] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[2] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, Feb (2012), 281–305.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11, Jul (2010), 2079–2107.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[6] Choi et al. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 787–795.

[7] Olivier Corby and Catherine Faron Zucker. 2010. The KGRAM abstract machine for knowledge graph querying. In *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 1. IEEE, 338–341.

[8] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

[9] George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* 12, 1 (2010), 49–57.

[10] Oana Frunza, Diana Inkpen, and Thomas Tran. 2011. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE transactions on knowledge and data engineering* 23, 6 (2011), 801–814.

[11] Raphaël Gazzotti, Catherine Faron Zucker, Fabien Gandon, Virginie Lacroix-Hugues, and David Darmon. 2019. Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction. In *ESWC 2019 - The 16th Extended Semantic Web Conference (Lecture Notes in Computer Science)*, Vol. 11503. Portorož, Slovenia, 116–130. https://doi.org/10.1007/978-3-030-21348-0_8

[12] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898* (2019).

[13] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 24, 1 (2017), 198–208.

[14] Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70.

[15] V Lacroix-Hugues, David Darmon, C Pradier, and Pascal Staccini. 2017. Creation of the First French Database in Primary Care Using the ICPC2: Feasibility Study. *Studies in health technology and informatics* 245 (2017), 462–466.

[16] Peter McCullagh and John A Nelder. 1989. *Generalized linear models*. Vol. 37. CRC press.

[17] Fco Javier Ordónez, Paula de Toledo, and Araceli Sanchis. 2013. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors* 13, 5 (2013), 5460–5477.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[20] Alberto G Salguero, Macarena Espinilla, Pablo Delatorre, and Javier Medina. 2018. Using Ontologies for the Online Recognition of Activities of Daily Living. *Sensors* 18, 4 (2018), 1202.

[21] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio, and Clement Jonquet. 2018. SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics* 19, 1 (2018), 405.

[22] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[23] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research* 39, suppl_2 (2011), W541–W545.

6 APPENDIX

ζ with the logistic regression algorithm (*LR*) uses the following parameters:

- Fold 1: 'C': 0.056049240151690681, 'penalty': 'l2'.
- Fold 2: 'C': 0.83617364781543058, 'penalty': 'l2'.
- Fold 3: 'C': 0.078134513655501683, 'penalty': 'l2'.
- Fold 4: 'C': 0.070037689307546724, 'penalty': 'l2'.
- Fold 5: 'C': 0.030094071461144355, 'penalty': 'l2'.
- Fold 6: 'C': 0.19901721018094651, 'penalty': 'l2'
- Fold 7: 'C': 0.16012788113832127, 'penalty': 'l2'.
- Fold 8: 'C': 0.067362109991791305, 'penalty': 'l2'.
- Fold 9: 'C': 0.034161307706627134, 'penalty': 'l2'.
- Fold 10: 'C': 0.055643396004174048, 'penalty': 'l2'.