Supervised Topic Models for Diagnosis Code Assignment to Discharge Summaries

Mohamed Dermouche, Julien Velcin, Rémi Flicoteaux, Sylvie Chevret, Namik Taright

Supervised Topic Models for Diagnosis Code Assignment to Discharge Summaries

Mohamed Dermouche[1,2], Julien Velcin[3], Rémi Flicoteaux[1,2,4], Sylvie Chevret[1,2,4], and Namik Taright[1,5]

[1] INSERM, U1153 Epidemiology and Biostatistics Sorbonne Paris Cité Research Center (CRESS), ECSTRA team, Paris, F-75010 France
mohamed.dermouche@inserm.fr, remi.flicoteaux@aphp.fr, sylvie.chevret@univ-paris-diderot.fr, namik.taright@aphp.fr
[2] Paris Diderot University, France
[3] Université de Lyon (ERIC LYON 2), France
julien.velcin@univ-lyon2.fr
[4] Saint-Louis Hospital, AP-HP, Paris, F-75010 France
[5] SIMAP/DOMU/AP-HP, Paris, F-75004 France

Abstract. Mining medical data has significantly gained interest in the recent years thanks to the advances in data mining and machine learning fields. In this work, we focus on a challenging issue in medical data mining: automatic diagnosis code assignment to discharge summaries, i.e., characterizing patient's hospital stay (diseases, symptoms, treatments, etc.) with a set of codes usually derived from the International Classification of Diseases (ICD). We cast the problem as a machine learning task and we experiment some recent approaches based on the probabilistic topic models. We demonstrate the efficiency of these models in terms of high predictive scores and ease of result interpretation. As such, we show how topic models enable gaining insights into this field and provide new research opportunities for possible improvements.

Keywords: ICD code assignment, topic models, machine learning, natural language processing, text categorization, text mining

1 Introduction

Health information systems are used at a large scale in the healthcare institutions and hospitals for various tasks, such as medical record management, medical prescription, and billing. As a result, increasing large volumes of healthcare data are regularly generated in the form of Electronic Medical Records (EMR). In this regard, textual data has a prominent place. Free text is actually a suitable form to describe a wide range of data related to patient's care including medical history, personal statistics, admission diagnosis, patient-caregiver exchange, etc. However, despite of being an abundant and valuable resource, only low quantities of these data are actually used for specific mining tasks, e.g., [14, 15].

One major issue that can be approached by capitalizing on the routinely generated textual data is the automation of diagnosis code assignment to medical

notes [4, 5, 9–11, 13–15, 17, 21]. The task involves characterizing patient's hospital stay (symptoms, diagnoses, treatments, etc.) by a small number of codes, usually derived from the International Classification of Diseases (ICD). Diagnosis codes provide a fast and easy understanding of patient's state evolution. The same codes are used as billing elements by the health insurance systems. Because of its importance, the task of code assignment is often performed manually by professional coders. However, manual coding is tedious and time-consuming: on average the coders spend about five minutes identifying only single code.

The main goal of this paper is to explore a new approach to automatic code assignment, based on probabilistic topic models. This approach has shown excellent performance in various text mining tasks, such as topic discovery, information retrieval, and sentiment analysis [8, 20]. Moreover, topic models provide a natural way to error analysis based on topic description, for example using discriminant words. Though, apart from some sparse work, the application of topic models for medical text mining purposes remains relatively less explored than in the other fields [8, 13, 20].

In this work, we experiment some recently-proposed supervised topic models in the task of automatic code assignment to medical discharge summaries. Our contributions can be summarized as follows:

1. New benchmark data: we create two french datasets with discharge summaries and manually associated ICD codes.
2. New learning models: we experiment some recent probabilistic topic models in a supervised fashion [1, 16].

Our evaluation setup allows a fair comparison of machine learning models in a more refined way than the traditional measures. Both code and data will be available for the community after anonymization.

A brief introduction to the ICD in given in Section 2. Then, an overview of prior work for diagnosis assignment is given in Section 3. Experiments (methods, data, and evaluation framework) are described in Section 4. Results and discussion are given in Section 5. Finally, the paper is concluded in Section 6.

2 International Classification of Diseases

According to the World Health Organization (WHO), the International Classification of Diseases (ICD) is the "standard diagnostic tool for epidemiology, health management and clinical purposes"[6]. This mainly includes diseases, but also symptoms, signs, procedures, and other content related to diseases. There exist a separate classification per language, that is regularly revised by the WHO. Currently, the latest revision for English is ICD10 whereas for French it is called CIM10 (*Classification Internationale de Maladies*). However, ICD9 is the most widely-used classification for diagnosis coding, in particular ICD9-Clinical Modification (ICD9-CM) as it allows comparability and use of mortality and morbidity data.

[6] `http://www.who.int/classifications/icd/`

Table 1. ICD code examples from CIM10 (top) and ICD9 (bottom).

ICD	Language	Code	Label
CIM10	French	C83.7	*Lymphome de Burkitt* (Burkitt's lymphoma)
		C88.0	*Macroglobulinmie de Waldenström* (Waldenström's macroglobulinemia)
		D30.0	*Tumeur bénigne du rein* (benign kidney tumor)
ICD9	English	198.3	Secondary malignant neoplasm of brain and spinal cord
		414.01	Coronary atherosclerosis of native coronary artery
		V34.01	Other multiple birth (three or more), mates all liveborn, delivered by cesarean section

For this paper, we use ICD9-CM (that we call ICD9) and CIM10 classifications for English and French respectively. In ICD9, diagnosis codes are 3-5 characters. The first character is numeric or alpha while characters 2-5 are numeric. In CIM10, diagnosis codes are far to 6 characters. The first character is always alpha and designs a high-level category, while the remaining are numeric. Table 1 shows some examples from the codes used in this paper. On the other hand, ICD can be structured in a tree hierarchy with edges representing "is-a" relationship between a parent code and its children. More details about ICD can be found on the WHO website[6].

3 Related Work

The problem of diagnosis code assignment has been studied from both perspectives of machine learning and computational linguistics, leading to a number of significant works. A bulk of these works have been published with the Computational Medicine Center's 2007 medical NLP challenge involving ICD code assignment to radiology reports [15]. With 45 distinct codes, the best F-score from the challenge was 89% while the average F-score was 77%. Note that all these works used multi-label classification: a document is assigned to one or more diagnoses (which is outside the scope of this paper). Some of them focused on using the ICD hierarchy to improve classification accuracy [14, 21].

In [5, 21], statistical classifiers were learnt based on a bag-of-words representation. The works in [5] used BoosTexter, a boosting-like technique based on a weak classifier, to learn a set of classification rules. The best achieved F-scores were around 84%. In [21], a simple classifier was learnt based on the presence/absence of UMLS terms[7]. The achieved scores were around 86%.

Besides this challenge, there also were some significant work such as [9, 13, 14, 17]. In [9], both SVM and Ridge Regression classifiers have achieved a score of 68% on a dataset with 2,618 distinct codes and a large number of learning documents (nearly 100,000). In [14], SVM classifier has been tested under a flat and a hierarchical setting. On a dataset with 5,030 distinct codes, the achieved F-scores were around 27% under flat setting and 39% under hierarchical setting. In [17], the k-NN classifier has been tested on a French corpus of medical

[7] `https://www.nlm.nih.gov/research/umls/`

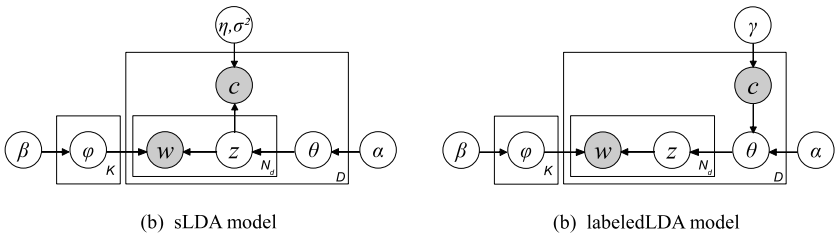
documents with more 10,000 distinct codes. The algorithm achieved about 74% precision score but very low recall levels.

The closest work to ours is by Perotte et al. [13]. The authors proposed a hierarchically-supervised topic model (HSLDA) combining LDA model [2] with the knowledge from ICD structure. The hierarchical structure of ICD codes was taken into account during the topic learning step. For this to happen, the final predicted code was constrained to derive from one branch of the tree (a code could not be assigned to a document if its parent were not). HSLDA have been tested on a dataset with 7,298 distinct codes, where it performed about 5% better than the non-hierarchical sLDA model [1]. Unfortunately, HSLDA source code is not publicly available which prevents us from including it in this study.

4 Experiments

4.1 Methods

We choose three traditional machine learning models: an example-based model (Decision Tree), a probabilistic model (Naive Bayes), and a kernel-based model (Support Vector Machines). We put these models against two others from the topic model family: sLDA [1] and labeledLDA [16]. To carry out the experiments, we rely on R rpart package that implements an efficient Decision Tree (DT) algorithm based on information-gain ratio as a splitting criterion [19]. Similarly, we use e1071 package to perform Naive Bayes (NB) and Support Vector Machines (SVM) classifiers [12]. For the latter one, the best performance is obtained with a linear kernel while all the remaining parameters are left to their default values.



(b) sLDA model (b) labeledLDA model

Fig. 1. Plate notation of (a) sLDA and (b) labeledLDA topic models. In labeledLDA, the topics (variable θ) are directly influenced by document's classes (variable c).

sLDA and labeledLDA are both based of the well-known LDA topic model [2]. Both models implement supervised learning based on the hidden topic structures (latent variables). In fact, the topic modeling process can be assimilated to a fuzzy word clustering where the goal is to build semantically coherent clusters [2]. sLDA and labeledLDA rely on slightly different structures (see Fig. 1). In both cases, the supervised part is implemented through a response variable, depicted by the letter c, that gives the predicted class modality (here the ICD codes).

Despite their apparent similarities, sLDA and labeledLDA differ mainly on two key points:

- Response formulation: in sLDA the response variable is derived from a Gaussian, while it follows a multinomial in labeledLDA. As such, labeledLDA would be more flexible and better fit with a multi-label classification, which is outside the scope of this paper.
- Topic supervision: in sLDA the response is calculated from empirical (learnt) topic distributions. For this, sLDA relies on a generalized linear model that maps the multinomial topic-document associations into a categorical response. In contrast, labeledLDA allows the knowledge from document's classes influencing topic construction. Thus, documents from the same classes are more likely to be linked with the same topics.

The parameters of sLDA and labeledLDA are fixed empirically in such a way to maximize the predictive scores on a held-out (test) sample[8]. For both models, the number of topics K is set to the number of codes. For sLDA: $\alpha = 0.01$. For labeledLDA, $\alpha = 0.005, \beta = 0.07$. The remaining hyperparameters are learnt from data. In addition, to maintain low running time, the number of iterations is set to 10,000 for sLDA and 50,000 for labeledLDA.

4.2 Datasets

As a response to the need for benchmark datasets pointed out in the literature [14], we created two datasets by gathering discharge summaries from Saint-Louis university teaching hospital[9]: URO-FR and HEMATO-FR. These datasets were built by taking all the discharge summaries collected within urology and hematology services respectively, between 2009 and 2014. Apart from filtering out rare codes (with less than 10 documents), we did not make any restriction regarding data quality, such as the presence of noise and typos. The point was to create real-life issued data with more challenging analysis problems for the algorithms. In this work, we only focus on the primary diagnosis (the reason the patient came to therapy) to deal with a single-label classification problem. We leave secondary codes, including aftercare codes, to a future work that will rely on multi-label classification.

The three datasets are highly imbalanced: about 70% of documents are assigned to 20% of codes. Once again, our goal is to experiment the models within the challenging real-world setup. Therefore, we choose to maintain the original imbalanced document distribution.

The third dataset MIMIC-EN is a subset of MIMIC-II Physiology database [18] using the following PostgreSQL query:

[8] Source codes from: `http://www.cs.cmu.edu/~chongw/slda/` (sLDA) and `https://github.com/myleott/JGibbLabeledLDA/` (labeledLDA).

[9] http://hopital-saintlouis.aphp.fr/

Table 2. Dataset description.

Dataset	ICD version	Lang.	#docs.	#unique words	#codes	Avg. #words /doc.	Avg. #docs. /code
URO-FR	CIM10	French	4 690	11 143	60	46	78
HEMATO-FR	CIM10	French	3 720	13 371	30	76	124
MIMIC-EN	ICD9	English	7 956	12 951	252	59	32

```
SELECT (subject_id, hadm_id, code, text)  FROM mimic2v26.icd9  JOIN
mimic2v26.noteevents USING (subject_id, hadm_id) WHERE (sequence='1'
AND category='DISCHARGE_SUMMARY' AND LENGTH(text) > 50);
```

The exact dataset used in this paper was obtained when discarding rare codes (a minimum of 15 documents has been chosen to make a trade-off between the total number of codes and the number of documents per code).

In order to mitigate the effects of high dimensionality, we systematically make the following text preprocessing:

1. Stemming
2. Filtering out the words occurring in less than 2 documents (3 for MIMIC-EN dataset because of its large size) or more than 300 documents
3. Removing stopwords and numerics.

The preprocessed text documents are then mapped into a bag-of-words representation where the words (unigrams) are weighted according to their presence/absence in the document. All the models are based on this representation. Table 2 gives an overview of the preprocessed datasets.

4.3 Evaluation framework

Previous works in code assignment have mainly relied on automatic evaluation measures from information retrieval fields, specifically precision, recall, and F-score. The standard version of these measures, yet widely-used for many predictive tasks, is restrictive in that it only considers a single model response. For this paper, we suggest to use a more flexible version, called F_k-score, in order to take into account uncertainty of the predictive models. To this end, the evaluation of a given result is performed by considering all of the k returned classes rather than one single class, as in [7, 17]. This choice is motivated by the following observations:

– All of the models tested here return a set of ranked labels (NB, sLDA, and labeledLDA) or can easily be adapted to do so (DT and SVM) [3]. Our evaluation enables retrieving the correct class in case it were not ranked first. This is particularly useful when some of the returned labels are ranked equally.

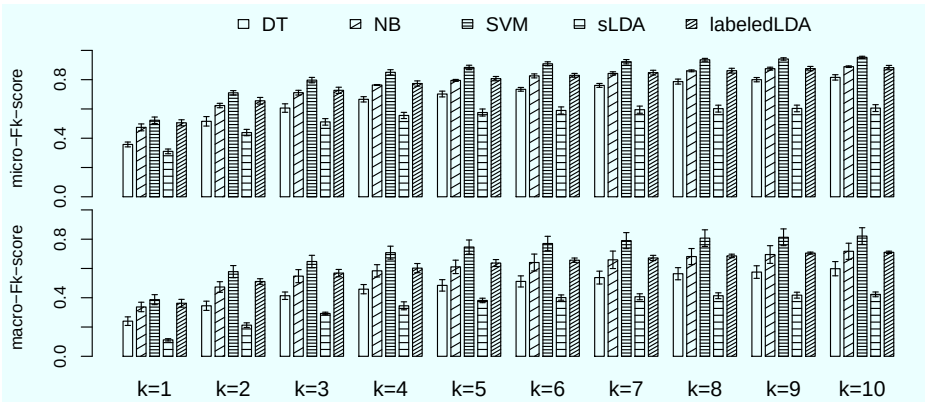


Fig. 2. Performance scores obtained on URO-FR dataset (60 codes).

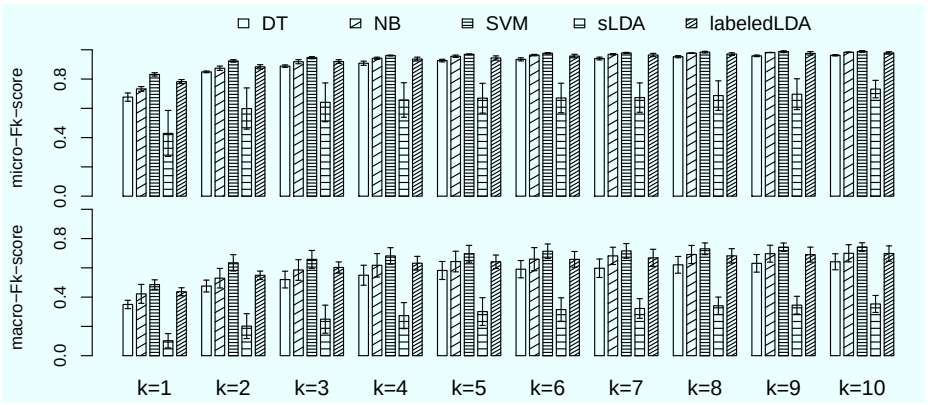


Fig. 3. Performance scores obtained on HEMATO-FR dataset (30 codes).

– In practice, it is more prudent to make the task humanly-supervised rather than fully automatic. In this regard, a set of best ranked codes is returned by the model, from where the coder selects the appropriate ones.

F_k-score is calculated similarly to the standard F-score except that the correct class is fetched among the k most probable classes returned by the model. That is, if the correct class is present within these classes, it is returned instead of the the most probable class.

5 Results and Analysis

Figures 2, 3, and 4 show the results from the three datasets. These are described in terms of micro (weighted average) and macro (average) F_k-scores, for k ranging from 1 to 10. The results are obtained based on a 10-fold cross validation.

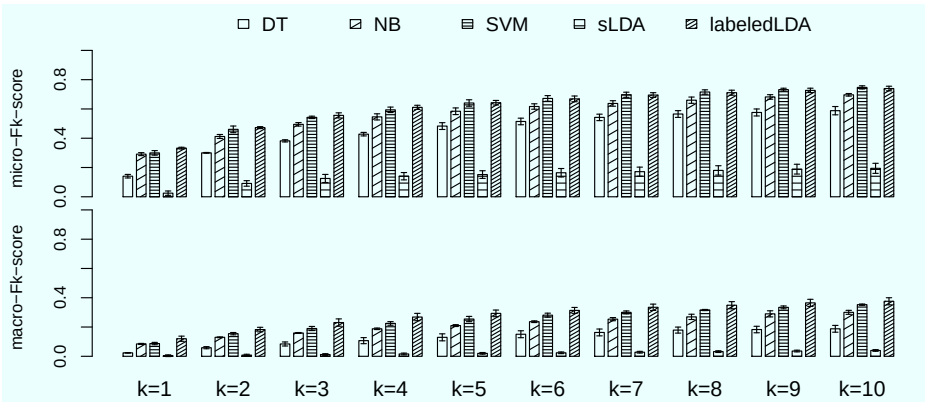


Fig. 4. Performance scores obtained on MIMIC-EN dataset (252 codes).

Table 3. F_k-scores for SVM and labeledLDA with $k \in \{1, 5, 10\}$.

Dataset	Model	micro-F_1-score	macro-F_1-score	micro-F_5-score	macro-F_5-score	micro-F_{10}-score	macro-F_{10}-score
URO-FR	SVM	**0.52 ±0.02**	**0.39 ±0.03**	**0.89 ±0.02**	**0.75 ±0.03**	**0.95 ±0.01**	**0.82 ±0.01**
	labeledLDA	0.51 ±0.02	0.36 ±0.03	0.81 ±0.02	0.65 ±0.03	0.88 ±0.02	0.71 ±0.01
HEMATO-FR	SVM	**0.83 ±0.01**	**0.49 ±0.03**	**0.95 ±0.01**	**0.69 ±0.02**	**0.99 ±0.01**	**0.74 ±0.03**
	labeledLDA	0.78 ±0.01	0.44 ±0.03	0.92 ±0.01	0.64 ±0.02	0.98 ±0.01	0.70 ±0.05
MIMIC-EN	SVM	0.30 ±0.02	0.09 ±0.01	0.61 ±0.01	0.24 ±0.02	**0.75 ±0.01**	0.35 ±0.02
	labeledLDA	**0.33 ±0.01**	**0.12 ±0.02**	**0.62 ±0.02**	**0.30 ±0.02**	0.74 ±0.02	**0.38 ±0.02**

Error bars give the standard deviations. In addition, Table 3 offers the exact scores for SVM and labeledLDA with $k \in \{1, 5, 10\}$. In the following, we discuss these results from three perspectives: (i) overall performance, (ii) performance w.r.t. k, and (iii) a comparison of topic models sLDA and labeledLDA.

Overall performance: as can be seen from these figures, SVM and labeledLDA always yield the best result compared to the other models. On URO-FR and HEMATO-FR, SVM has the best scores while on MIMIC-EN labaledLDA comes first. Based on a t-test throughout all the datasets, no evidence of any statistical difference could be observed between SVM and labeledLDA (p-value > 0.05). NB generally arrives third, followed by DT, then sLDA that performs comparably poor in this task, specifically on MIMIC-EN where the number of codes is large. The same trend can be observed with both micro and macro averaged F-scores.

Performance w.r.t. k: better results are achieved when the value of k grows up, giving more chance to the less probable codes to be selected. By averaging over all models and datasets, the gain in micro-F_k-score is equal to 14% when k increases from 1 to 2. It is equal to 2% when k increases from 5 to 6, and 1% when it increases from 9 to 10. To assess the statistical significance, we perform a paired t-test on the micro-F_k-scores obtained from MIMIC-EN dataset. The p-values result from comparing the means of micro-F_{k+1}-scores vs. micro-F_k-scores.

Table 4. Examples of codes and associated topics from URO-FR (top), HEMATO-FR (middle), and MIMIC-EN (bottom) datasets extracted with labeledLDA model.

C61: *Tumeur maligne de la prostate* (Prostate cancer)	**N39.3**: *Incontinence urinaire d'effort* (Stress urinary incontinence)	**Z52.4**: *Donneur de rein* (Kidney donor)	**N30.0**: *Cystite aiguë* (Acute cystitis)	**S30.2**: *Contusion des organes génitaux externes* (Congestion of the external genitalia)
prostatectomie[5]	*incontinent*	*prelèvement* (sample)	*pontage* (bypass)	*observer* (watch)
radical	*bandelette* (band)	*faveur* (favour)	*arterielle* (arterial)	*hospitalisé*(inpatient)
laparotomie (laparotomy)	*effort* (stress)	*manuel* (hand-operated)	*Ditropan*	*med* (medical)
score	*trans-obturatrice*[5]	*artère* (artery)	*post-mictionnel*[5]	*externe* (lateral)
lobe (lobus)	*urodynamique*[5]	*assisté* (assisted)	*Kardegic*	*motif* (cause)
mini	*toux* (cough)	*DFG* (GFR)	*diurne* (diurnal)	*chir* (surgery)
capsulaire (capsular)	*bud* (urodynam. test)	*laparoscopique*[5]	*surtout* (especially)	*ATCD* (med. history)
élevé (high)	*rééducation*[5]	*contre* (against)	*fonctionnel*(functional)	*clinique-uro*[5]
extension	*urgenturie*[5]	*apparenté* (related)	*impériosité* (urge)	*fam* (familial)
curatif (curative)	*position*	*min* (minute)	*hypertension*[5]	*suggérer* (suggest)
#documents=356	#documents=47	#documents=39	#documents=16	#documents=18
F_1-score=0.68	F_1-score=0.83	F_1-score=0.96	F_1-score=0.00	F_1-score=0.22

C81.9: *Lymphome de Hodgkin* (Hodgkin's lymphoma)	**C88.0**: *Macroglobulinmie de Waldenström* (Waldenström's macroglobulinemia)	**D46.2**: *Anémie réfractaire avec excès de blastes* (refractory anemia with excess of blasts)	**C83.0**: *Lymphome à petites cellules B* (small B-cell lymphoma)	**E85.3**: *Amylose généralisée secondaire* (secondary generalized amyloidosis)
Hodgkin	*Waldenström*	*senior*	*critère (criterion)*	*amylose*
ABVD	*IgM*	*multirésistant*(resistant)	*participer(participate)*	*troponine* (troponin)
IVOX	*lymphoplasmocytaire*[10]	*remise* (redelivery)	*accepter (accept)*	*formule* (formula)
classique (classical)	*macroglobulinémie*[5]	*blaste* (blast)	*consentement(consent)*	*proBNP*
panoramique(panoramic)	*monoclonal*	*AREB* (RAEB)	*aborder (approach)*	*VCD*
escalade (escalation)	*béta* (beta)	*leuco*	*attendu (expected)*	*évolution* (evolution)
étoposide (etoposide)	*créatininémie*[5]	*Vidaza*	*logistique (logistics)*	*dosage* (dose)
BEAM	*sup* (increased)	*myélodysplasique*[5]	*version*	*arriver* (reach)
SPI (IPS)	*stabilité* (stability)	*BHC*	*objectif (goal)*	*immunochimique*[5]
nodulaire (nodular)	*cérébral* (cerebral)	*mgX* (m.g.)	*contrainte(constraint)*	*physique* (physical)
#documents=168	#documents=72	#documents=37	#documents=38	#documents=85
F_1-score=0.75	F_1-score=0.74	F_1-score=0.78	F_1-score=0.38	F_1-score=0.34

157.0: Malignant neoplasm of pancreas	**278.01**: Morbid obesity	**430**: Subarachnoid hemorrhage	**038.0**: Streptococcal septicemia	**998.12**: Hematoma complicating a procedure
duct	morbid	coil	vegetation	FFP
painless	roxicet	vasospasm	biliary	tube
biliary	elixir	nimodipine	streptococcus	yellow
bile	roux-en-y	fluent	surveillance	soften
whipple	crush	downgoing	endocardial	layer
CBD	laparoscopic	cistern	enterococcus	fiber
ERCP	actigall	angio	ductal	etc
endoscopic	bloated	pronation	cellular	colitis
duodenum	pill	sah	travel	everyday
cholangiopancreatography	program	EOM	medial	sleep
#documents=13	#documents=13	#documents=89	#documents=26	#documents=9
F_1-score=1.00	F_1-score=1.00	F_1-score=0.69	F_1-score=0.00	F_1-score=0.00

[5] Term translation: *clinique-uro*: clinical-urological, *créatininémie*: creatininemia, *immunochimique*: immunochemical, *laparoscopique*: laparoscopic, *lymphoplasmocytaire*: lymphoplasmocytic, *macroglobulinémie*: macroglobulinemia, *myélodysplasique*: myelodysplastic, *post-mictionnel*: post-void, *prostatectomie*: prostatectomy, *rééducation*: reeducation, *trans-obturatrice*: transobturator, *urodynamique*: urodynamics, *urgenterie*: urge incontinence.

The difference is highly significant when k increases from 1 to 2 or from 2 to 3 (p-value $< 10^{-6}$). In contrast, the difference is comparably much less significant for the greater values of k (p-value $> 10^{-3}$).

sLDA vs labeledLDA: sLDA clearly achieves lower scores than the other topic model labeledLDA. For $k = 1$, the difference in micro-F_k-score is equal to 20% on URO-FR, 35% on HEMTAO-FR, and 30% on MIMIC-EN (see Table 3). We believe that this great difference in performance is due to the intrinsic difference in model structure. In labeledLDA, the knowledge from document's classes directly influences the topic construction. As such, documents from the same class (diagnosis code) are more likely to link to the same topics, which helps building more "diagnosis-based" topics. This feature, not shared by sLDA, is graphically depicted by the direction of the edge linking the variables c and θ (see Fig. 1).

In addition to these quantitative results, we show the top 10 words characterizing the topics obtained with labeledLDA in Table 4. Each code is associated with the most likely topic based on empirical distributions [16]. We choose three examples from the best predicted scores and two examples with poorly predicted ones. The underlined words are manually annotated by medical experts as being semantically and clearly related to the associated diagnosis. Medical experts agree that these results are very informative. Most of the diagnoses are easily recognizable from their characterizing words. Moreover, a post hoc analysis of these results leads to the following observations:

- Topic's coherence is generally correlated to good predictive scores, as with the codes C81.9, C88.0, C91.1, N20.0 from the French data. Conversely, the codes with less coherent and/or mixed topics have poor predictive scores, such as C81.9, C83.0, N15.1, N20.1, N30.0. This observation may help explaining why certain codes are so easily-predicted by the model whereas others are not. In this regard, it is legitimate to believe that improving the topic's quality would lead to improve the predictive scores.
- A large number of codes can be characterized with medical concepts (n-grams, phrases) rather than single words, for example "arterial tension", "urinary tract infection", "blood test", etc. This observation should motivate the inclusion of medical concepts, either extracted statistically or based on medical ontology, into the vocabulary.

6 Conclusion

The work described in this paper is an example application of machine learning models to a real-world problem: diagnosis code assignment to discharge summaries. The models that we have chosen for experiments are issued from both classical machine learning research (DT, NB, and SVM) and modern NLP approaches (sLDA and labeledLDA). Despite the achieved results that are quite encouraging, the task would not allow a fully automatic coding because of the significant error rates. For example, based on labeledLDA model, 22% documents

from hematology service would be miscoded. The rate rises to 51% on urology and 65% on intensive care medicine. A thorough analysis of prediction errors suggests that data quality (such as size, coverage, and specificity) is crucial for the success of the task. The code distribution is also an important factor as the codes with small sample sizes are generally hard to predict.

After a thorough discussion with medical experts, we believe that the automatic part of the coding process is very useful but cannot dispense with human supervision to make the conclusive choice. Nevertheless, as it has been shown in Table 3, the error rates are dramatically reduced when allowing larger values of k. For example, with k=10, the error rates are reduced by 25–45 percentage points. In this way, the human coder can seek for the appropriate codes in a reduced space, which makes the coding task faster and easier.

Finally, in a semi-automatic approach the coder also has the ability to produce and express feedback for the learning algorithms. This can be done either by using the coder's choice among the proposed codes, or in a more specific way by asking the user to express a prior (e.g., "word-code" relation). Beyond the natural extension of this work to a multi-label task (including both primary and secondary codes), a challenging future work would be how to efficiently include user feedback, and more generally any type of prior knowledge. Conveniently, topic models are highly flexible for such purposes (see seededLDA model in [6]). Convinced by the utility of using prior knowledge and motivated by the promising results achieved with labeledLDA model, we consider extending our work to the embedding of this knowledge into labeledLDA model.

References

1. David M. Blei and Jon D. Mcauliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS'07)*, pages 121–128, Vancouver, Canada, 2007. Curran Associates, Inc.
2. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
3. Ricardo Cerri, André Carlos P. L. F. De Carvalho, and Alex A. Freitas. Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. *Intelligent Data Analysis*, 15(6):861–887, 2011.
4. Richárd Farkas and György Szarvas. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics*, 9 Suppl 3:S10, 2008.
5. Ira Goldstein, Anna Arzrumtsyan, and Ozlem Uzuner. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In *Proceedings of AMIA Symposium (AMIA'07)*, pages 279–83, 2007.
6. Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating Lexical Priors into Topic Models. In *Proceedings of the European Chapter of the ACL (EACL'12)*, pages 204–213, Avignon, France, 2012. ACL.
7. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS'12)*, pages 1106–1114, Lake Tahoe, NV, USA, 2012. NIPS.
8. Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(6):1134–1145, June 2012.

9. Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. Large Scale Diagnostic Code Classification for Medical Patient Records. In *Proceeding sof the International Joint Conference on Natural Language Processing (IJCNLP'08)*, pages 877–882, Hyderabad, India, 2008. ACL.
10. Julia Medori and Cédrick Fairon. Machine learning and features selection for semi-automatic ICD-9-CM encoding. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi'10)*, pages 84–89, Los Angeles, CA, USA, 2010. ACL.
11. Elisabeth Metais, Didier Nakache, and Jean-François Timsit. Automatic Classification of Medical Reports , the CIREA project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics (TELE-INFO'06)*, pages 354–359, Istanbul, Turkey, 2006. WSEAS.
12. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group, 2015.
13. Adler Perotte, Nicholas Bartlett, Frank Wood, and Noemie Elhadad. Hierarchically Supervised Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems (NIPS'11)*, pages 2609–2617, Granada, Spain, 2011.
14. Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association (JAMIA)*, 21(2):231–237, 2014.
15. John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A Shared Task Involving Multi-label Classification of Clinical Free Text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP'07)*, pages 97–104, Prague, Czech Republic, 2007. ACL.
16. Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, number August, pages 248–256, Singapore, Singapore, 2009. ACL.
17. Patrick Ruch, Julien Gobeilla, Imad Tbahritia, and Antoine Geissbühlera. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. In *Proccedings of the AMIA Symposium (AMIA'08)*, pages 636–40, Washington D.C., USA, 2008.
18. Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960, 2011.
19. Terry Therneau, Beth Atkinson, and Brian Ripley. rpart: Recursive Partitioning and Regression Trees, 2015.
20. Xing Yi and James Allan. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*, pages 29–41, Toulouse, France, 2009. Springer-Verlag.
21. Yitao Zhang. A hierarchical approach to encoding medical concepts for clinical notes. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop (HLT-SRWS'08)*, pages 67–72, Columbus, OH, USA, 2008. ACL.