ECSTRA-INSERM @ CLEF eHealth2016-task 2: ICD10 Code Extraction from Death Certificates

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

[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] AP-HP, Paris, F-75004 France
vincent.looten@aphp.fr
[4] Saint-Louis Hospital, AP-HP, Paris, F-75010 France
[5] Université de Lyon (ERIC Lyon 2), France
julien.velcin@univ-lyon2.fr

Abstract. This paper describes the participation of ECSTRA-INSERM team at CLEF eHealth 2016, task 2.C. The task involves extracting ICD10 codes from death certificates, mainly described with short plain texts. We cast the task as a machine learning problem involving the prediction of the ICD10 codes (categorical variable) from the raw text transformed into a bag-of-words matrix. We rely on probabilistic topic models that we evaluate against classical classifiers such as SVM and Naive Bayes. We demonstrate the effectiveness of topic models for this task in terms of prediction accuracy and result interpretation.

Keywords: ICD10 code assignment, cause of death extraction, topic models, machine learning, natural language processing, text mining

1 Introduction

Completing death certificates is a routine task in hospitals and healthcare institutions. In France, the death certificates are produced by physicians and transmitted to the French Epidemiological Center for the Causes of Death (CépiDC)[6]. Beyond the administrative and personal information, the death certificates usually contain a free-text description of the cause(s) of death. To monitor population's health, free texts are converted by the CépiDC into formal standardized codes, usually derived from the International Classification of Diseases (ICD) taxonomy. These codes also serve as a basis for mortality and epidemiology studies.

The ICD taxonomy covers a wide range of diseases, symptoms, signs, procedures, and other content related to diseases[7]. The World Health Organization

[6] http://www.cepidc.inserm.fr/
[7] http://www.who.int/classifications/icd/en/

issues separate versions of ICD per language/country. In this paper, we use the French release of ICD, which is now at its 10^{th} revision (called ICD10). It covers more than 20,000 codes including diagnoses and procedures, but only a subset of theses codes can be causes of death. An example is provided in Table 1.

Table 1. Example causes of death from the French ICD10 taxonomy.

C384	*Tumeur maligne de la plèvre* (malignant neoplasm of pleura)
C450	*Mésothéliome pleural* (mesothelioma of pleura)
E274	*Insuffisance surrénale aiguë* (unspecified adrenocortical insufficiency)
I678	*Epilepsie vasculaire* (vascular epilepsy)
X74	*Lésion auto-infligée par décharge d'armes à feu* (intentional self-harm by firearm and gun discharge)

Requiring manual work and expertise, the task of ICD10 code extraction from text is quite time-consuming because the ICD10 taxonomy contains thousands of possible causes of death. Within the CLEF eHealth 2016, the task 2.C focuses on the problem of automatic extraction of the causes of death from the textual description [8, 10]. The task may be approached either from a machine learning perspective (supervised classification) or a natural language processing perspective by using syntactic and/or semantic decision rules. Both approaches aim at automating the ICD10 code extraction from death certificates.

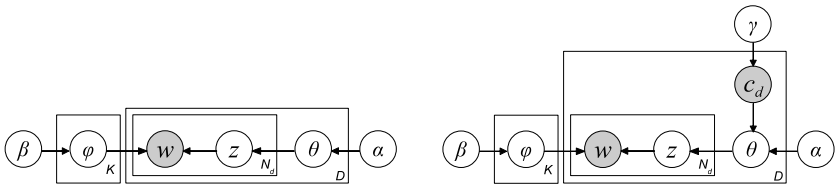
In this paper, we describe our system to automatic cause-of-death extraction following the first approach. Concerning the used methods, we mainly focus on probabilistic topic models [3, 14] that we evaluate against traditional machine learning methods, like SVM and Naive Bayes, with respect to predictive accuracy and result interpretation. We show that topic models are competitive with traditional methods in terms of predictive accuracy. We also show that topic models offer more easily-interpreted results and allow to gain a better insight in the data analysis.

2 Methods

Topic models are probabilistic approaches to discovering hidden structures (commonly called topics) from text. Topic models have shown significant efficiency over baseline models for various language modeling and text mining tasks, like topic discovery [3], information retrieval [18], and text classification [1]. The general approach is to model the co-occurrence relation among words (observed variables) and build soft clusters of words characterizing the topics.

Latent Dirichlet Allocation (LDA) [3] is one of the most popular topic models completely built on the co-occurrence assumption. In LDA, the words that tend to co-occur in the same documents are more likely to characterize the same topic. Conversely, the words that rarely co-occur are likely to describe different topics.

In LDA, the document's words are captured using an observed variable w (see Figure 1.a) while the topics are inferred as a latent variable z using Bayes calculus. For inferring the latent variables, a number of optimization methods can be deployed, like Gibbs sampling, Expectation Maximization, etc. For more details on the optimization process, we refer the uninitiated reader to the tutorials in [6, 15].



(a) LDA model (b) LabeledLDA model

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

Relying solely on word co-occurrence, LDA model is fully unsupervised. In this work, we rely on a supervised extension of LDA called LabeledLDA [14]. Unlike LDA, LabeledLDA is supervised in that it learns, in addition to the hidden topics, a "response variable". For many tasks, LabeledLDA has shown competitive results compared to traditional machine learning models, like Naive Bayes and SVM.

To better explain how LabeledLDA works, we rely on the plate notation given in Figure 1.b. In the traditional machine learning methods, the words (variable w) directly influence the prediction decision. For example, in Naive Bayes classifier, the decision depends on the word frequencies per class. In LabeledLDA, The prediction decision depends on both document's words and topics. That is, the topic's proportions of the document to be classified are taken into account when computing the document-code probabilities. On the other hand, the topic extraction with LabeledLDA is influenced by the document's classes. Thus, the documents from the same class are likely to be linked to the same topics. This feature allows LabeledLDA to take advantage of the hidden structures of documents as well as the topic-code relations. For example, the code N189 (chronic kidney disease) is likely to describe a topic about loss in kidney function, that in turn can be described using the words "kidney", "renal", "failure", etc.

The efficiency of LabeledLDA for text classification has been proved in many tasks, including diagnosis code assignment to medical summaries [4, 11], which is very close to our task. In this work, we aim at experimenting LabeledLDA model for the specific task of death cause extraction. The main differences from the work in [4] are the following:

– The task is not the same, here the ICD10 codes correspond to the causes of death while in [4] the codes correspond to diagnoses.

– The number of possible codes is much more important (3,231 vs. 60 in [4]) which endows the task with more challenging analysis problems.
– The documents used here are much shorter (a document contains 3.6 words on average vs. 60 words in [4]).

The parameters of LabeledLDA are fixed empirically in such a way to maximize the predictive scores on a held-out dataset (20% documents randomly sampled from the whole dataset). As such, the number of topics K is equal to the number of codes (3,231), $\alpha = 0.005, \beta = 0.07$. The remaining hyperparameters are learnt from data. In addition, we perform two experiments by setting the number of iterations to 20,000 and 50,000 respectively.

The two other classifiers used in this work are SVM and Naive Bayes. SVM is a non-probabilistic binary classifier that maps the documents in such a way to maximize the gap between those from opposite classes. For our non-binary problem, we use the one-vs-one technique that consists of learning a separate classifier of each pair of classes then taking the class with the largest weight. Naive Bayes is a simple and widely-used probabilistic classifier based on the calculation of conditional probabilities (probability of words given the outcome class). The process is then easily inverted using Bayes theorem and word independence assumption. To run these methods, we rely on Python and "scikit-learn" package[8], specifically "BernoulliNB" and "LinearSVC" implementations. For SVM, we set the penalty parameter c to 1.0. The rest of parameters are left to default values.

3 Dataset

The CépiDC corpus has been created by the French Center for Epidemiology and Medical Causes of Death (CépiDC) specifically for the CLEF eHealth 2016 contest [10]. It is composed of separate train/test samples of death certificates. Only the textual description of the causes of death are available for analysis. CépiDC dataset is highly imbalanced: about 80% of documents are assigned to less than 20% of codes.

Also, in order to reduce dimensionality, we filter out frequent words (contained in more than 50% of documents) and rare words (contained in less than 3 documents). We also remove stopwords and numerics. The preprocessed text documents are then mapped into a bag-of-words representation where the words are weighted according to their presence/absence in the document (binary values). Table 2 gives an overview of the preprocessed CépiDC sample used for training.

4 Results and Discussion

The methods are evaluated and ranked based on micro-averaged F-score (harmonic mean of precision and recall weighted by the class size). LabeledLDA

[8] http://scikit-learn.org/

Table 2. CépiDC training dataset description

Lang.	#docs.	#unique words	#codes	Avg. #words /doc.	Avg. #docs. /code
French	266,807	9,332	3,231	3.6	9.1

model is put against other traditional machine learning methods: Naive Bayes and SVM. We have chosen these methods among others because they gave the best scores.

Table 3. Micro-averaged predictive scores obtained on CépiDC dataset. The methods are sorted by F-score.

Method	Precision	Recall	F-score
SVM, linear kernel	**88.16**	65.54	**75.19**
LabeledLDA, 50,000 iterations	81.93	66.69	73.53
LabeledLDA, 20,000 iterations	81.10	61.53	69.97
Multinomial Naive Bayes	59.44	79.95	68.18

The obtained results are given in Table 3. For the official evaluation, we have submitted a run of LabeledLDA with 20,000 iterations because the running time was too long. In Table 3, we also evaluate a run of LabeledLDA with 50,000 iterations. Our two official submissions are given in italic whereas the best scores are given in bold.

As can be seen, SVM achieves the best F-score compared to other methods, followed by LabeledLDA model. In terms of micro-averaged F-score, SVM achieves 75.19% while LabeledLDA arrives second with 73.53%. Compared to the other 6 participant systems, our LabeledLDA-based system was ranked fourth (based on 20,000 iterations). When rising the number of iterations to 50,000, LabeledLDA would be ranked second. Based on the official results from [10], the average score from all participants was 71.85% while the median was 69.97%.

Although the superiority of SVM over LabeledLDA in terms of precision score, LabeledLDA's most interesting advantage resides in offering more easily-interpreted results. In Table 4, we show the top-10 words characterizing topics extracted with LabeledLDA model. To complete these results, we have asked a physician to decide of the significance of topics by looking at the top-10 words globally. The physician has agreed that topics were extremely informative. Most of the causes of death represented here could easily be recognized from the associated topical words.

Table 4. Examples of codes and associated topics from LabeledLDA model.

Code	Label	Probable words
X74	*Lésion auto-infligée par décharge d'armes à feu* (intentional self-harm by firearm and gun discharge)	*arme* (gun), *feu* (fire), *plaie* (wound), *thoracique* (thoracic), *cranio* (cranio), *projectile* (bullet), *gros* (large), *calibre* (calibre), *cardiaque* (cardiac), *cérébral* (cerebral)
C450	*Mésothéliome pleural* (mesothelioma of pleura)	*pleural* (pleural), *mésothéliome* (mesothelioma), *malin* (malignant), *droit* (right), *mésothélium* (mesothelium), *gauche* (left), *métastatique* (metastatic), *terminal* (terminal), *plèvre* (pleura), *épithelioide* (epithelioid)
C384	*Tumeur maligne de la plèvre* (malignant neoplasm of pleura)	*sarcome* (sarcoma), *thoracique* (thoracic), *pulmonaire* (pulmonary), *multimétastatique* (multi metastatic), *artère* (artery), *récidive* (recurrence), *évolutif* (evolutive), *thorax* (thorax), *pariétal* (parietal), *paroi* (lining)
E274	*Insuffisance surrénale aiguë* (unspecified adrenocortical insufficiency)	*insuffisance* (failure), *surrénalienne* (adrenal), *aiguë* (acute), *chronique* (chronic), *diabète* (diabetes), *surrénale* (adrenal), *rénal* (renal), *insulino* (insulin), *requérant* (petitioner), *hypophysaire* (hypophyseal)
I678	*Epilepsie vasculaire* (vascular epilepsy)	*épilepsie* (epilepsy), *AVC* (stroke), *épileptique* (epileptic), *encéphalopathie* (encephalopathy), *alzheimer* (alzheimer), *evoluée* (evolved), *syndrôme* (syndrome), *séquelles* (aftermath), *post* (after), *maladie* (disease)

5 Related Work

The problem of ICD code extraction has been investigated from a larger perspective involving code assignment to various types of medical documents. The cause of death may be considered as a specific task. The majority of these works have focused on English documents. Number of them have been published with the Computational Medicine Center's 2007 medical NLP contest involving ICD code assignment to radiology reports [13]. The dataset contained 45 different diagnosis codes and each document could be labeled with one or more codes (multi-label task). [5] used BoosTexter, a boosting-like technique based on a weak classifier, to learn a set of classification rules. In [19], a simple classifier was learnt based on the presence/absence of medical concepts from UMLS ontology[9]. The best micro-averaged F-score achieved within the contest was about 89%.

Apart from this contest, in [9] both SVM and Ridge Regression classifiers have achieved a score of 68% on a dataset with 2,618 distinct codes and about 100,000 documents. In [12], SVM classifier has been tested considering both flat and hierarchical setting. The hierarchical setting relies on the tree structure of ICD codes to improve accuracy. On a dataset with 5,030 distinct codes, the achieved F-scores were about 27% under flat setting and about 39% under hierarchical setting.

In [4], LabeledLDA has been used to extract ICD codes from both English and French discharge summaries. Compared to SVM classifier, LabeledLDA achieved almost same performance. With 60 distinct codes, the micro-averaged F-score was about 52%. In [11], the authors proposed a hierarchically-supervised topic model (HSLDA) that combines traditional topic modeling with the ICD struc-

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

ture. The hierarchical structure of ICD codes has been taken into account during the topic learning step. Towards this end, the final predicted code was constrained to derive from a single branch of the tree: a code could not be assigned to the document if its parent in the tree were not. Compared to a non-hierarchical version [2], HSLDA performed about 5% better on a dataset with 7,298 distinct codes. The source code of HSLDA has not been made available, which prevented from experimenting it for this task.

6 Conclusion

In this paper, we have described our system used to extract ICD10 codes from death certificates within the CLEF 2016 eHealth, task 2.C. Our system is based on LabeledLDA model: a supervised topic model for topic discovery and document classification. Even if LabeledLDA does not outperform SVM classifier, LabeledLDA provides an explanation of the results (why such code for such document?) and allows a more in-depth understanding of the classification mechanisms. This feature is obviously a clear advantage over machine learning methods, like SVM that are usually based on complex mechanisms and consequently less suitable for human interaction.

As a promising future direction, we believe that the performance of LabeledLDA can be improved by integrating an "active learning" component. That is, the active learning allows capturing and integrating user's feedback into the learning process [16, 17]. As such, it will also be possible to focus on specific data, for example a subset of misclassified documents, chosen by the user based on her experience [7]. The high flexibility offered by topic models will allow for more in-depth error analysis and better understanding of data, which is more convenient for integrating user interaction.

References

1. David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, pages 127–134, Toronto, Canada, 2003. ACM.
2. 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.
3. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
4. Mohamed Dermouche, Julien Velcin, Rémi Flicoteaux, Sylvie Chevret, and Namik Taright. Supervised Topic Models for Diagnosis Code Assignment to Discharge Summaries. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'16)*, Konya, Turkey, 2016. Springer.
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. Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2005.
7. Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, 2014.
8. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. CLEF 2016 - 7th Conference and Labs of the Evaluation Forum. *Lecture Notes in Computer Science (LNCS)*, 2016.
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. Aurélie Névéol, Lorraine Goeuriot, Liadh Kelly, Kevin Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Grégoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum. Clinical information extraction at the CLEF eHealth Evaluation lab 2016. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, Évora, Portugal, 2016. CEUR-WS.
11. 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.
12. 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.
13. 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.
14. 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.
15. Philip Resnik and Eric Hardisty. Gibbs Sampling for the Uninitiated. Technical Report June, 2010.
16. Yi Yang, Doug Downey, and Jordan Boyd-graber. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*, number r, pages 308–317, Lisbon, Portugal, 2015. ACL.
17. Yi Yang, Shimei Pan, Jie Lu, Mercan Topkara, and Doug Downey. Incorporating User Input with Topic Modeling. In *CIKM 2014 Workshop on Interactive Mining for Big Data (ImBig'14)*, Shanghai, China, 2014. ACM.
18. 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.
19. 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.