Anonymization of Sensitive Information in Medical Health Records

Bhavna Saluja*, Gaurav Kumar*, João Sedoc, and Chris Callison-Burch

University of Pennsylvania, Philadelphia PA 19104, USA
{bsaluja,gauku,joao,ccb}@seas.upenn.edu

Abstract. Due to privacy constraints, clinical records with protected health information (PHI) cannot be directly shared. De-identification, i.e., the exhaustive removal, or replacement, of all mentioned PHI phrases has to be performed before making the clinical records available outside of hospitals. We have tried to identify PHI on medical records written in Spanish language. We applied two approaches for the anonymization of medical records in this paper. In the first approach, we gathered various token-level features and built a LinearSVC model which gave us F1 score of 0.861 on test data. In the other approach, we built a neural network involving an LSTM-CRF model which gave us a higher F1 score of 0.935 which is an improvement over the first approach.

Keywords: PHI · Neural Networks · LSTM-CRF · De-identification · Anonymization · Computational Linguistics · Privacy

1 Introduction

Clinical records with protected health information (PHI) cannot be directly shared as is, due to privacy constraints, making it particularly cumbersome to carry out NLP research in the medical domain. A necessary precondition for accessing clinical records outside of hospitals is their de-identification, i.e., the exhaustive removal, or replacement, of all mentioned PHI phrases.

PHI stands for Protected Health Information and is any information in a medical record that can be used to identify an individual, and that was created, used, or disclosed in the course of providing a health care service, such as a diagnosis or treatment. In other words, PHI is personally identifiable information in medical records, including conversations between doctors and nurses about treatment. PHI also includes billing information and any patient-identifiable information in a health insurance company's computer system. Some information that can be considered PHI are Names, Surnames, Addresses, Hospitals, Professions, Different types of locations (provinces, cities, towns,...), Billing information, Email, Phone records.

* Equal contribution. Listing is in random order.

2 Literature Review

In paper [3], the authors have presented a deep learning architecture that uses bi-directional long short-term memory networks (Bi-LSTMs) with variational dropouts and deep contextualized word embeddings while also using components such as traditional word embeddings (Glove), character LSTM embeddings and conditional random fields. The paper [1] aims to develop techniques and methods for semi-automated anonymization of medical record information. The paper proposes methods like utilization of database structure, dictionaries, heuristics and natural language processing for anonymizing patient records in general. Major challenges which are posited are the differences in identity markers (e.g. Dr. and Mrs.) and hyphenation patterns in Norwegian, unstructured text, no strictly enforced guidelines for how the data should be encoded. In paper [2], the authors have developed a de-identification model that can successfully remove personal health information (PHI) from discharge records to make them conform to the guidelines of the HIPAA. Authors have used feature set of 5 different categories - Word level features, Frequency Information, Offline Dictionaries, Contextual Information and Phrasal Information. Authors have trained three different classifiers that used three different contextual features and used a voting based mechanism to decide if the word belonged to NER. If any two classifiers have predicted the same label, the word is assigned that tag otherwise its not considered NER. In [4], the authors proposed a three step approach to extract personal information from medical records. First, they split the document into terms and extract local and external features. Then they built multiple independent classifiers from the features that are extracted. At last, they combine the results of independent classifiers to get final tags of the words. In paper [5], the authors have experimented using SVMs to recognize NER data. Authors have built feature set using various features such as token level features like orthographic features, length, POS tag, kind etc. and features like date, id, phone number etc. They have used ANNIE Web API to identify hospitals, people or locations etc.

3 Data Preparation

We created dictionary representations of train, dev and test datasets from the clinical records given in text files [6]. We represented each word as:

$$(word, NER\_Tag, start\_index, Spanish\_POSTAG)$$

For each clinical record, we store a list of sentences. Each sentence is further a list containing tuple representation of words in it. The records are then stored as (key, value) pairs in a dictionary where key is docId and value is the sentences in the form of list of word tuples which are dumped as pickle files. These pickles are being used as input to all our models. We also created vocabularies for words, tags and characters present in our dataset and prepared a numpy array which contains the embeddings for tokens using fastText Spanish embeddings [7].

The given data set was divided into training set, development set and test set and the distribution is shown in Table 1.

Table 1. Dataset distribution

Type	Size	Avg Number of Sentences	Avg Number of Words
Training Set	401	20	393
Development Set	193	20	383
Test set	156	20	424

Evaluation of the model is done by measuring F-score performance metric which is a widely used metric in the natural language processing literature, such as the evaluation of named entity recognition and word segmentation. The published papers mentioned in the literature review section have also used F-score to evaluate the performance of their models.

4 Approaches

4.1 LinearSVC

We have built a baseline model implementing linear SVM on our dataset focusing on token-level features such as inclusion of punctuations, uppercase or lowercase letters in the token, or is the word Roman, fax-related features, etc. For each word we look at the window [-1,0,1,2] and create feature vector including the features of these words in the context of the target word. We have a total of 401 training documents, 193 dev documents and 156 test documents and we tokenized the files into sentences whose counts are mentioned as follows:

Train sentences = 8300 (Total 401 docs)
Dev sentences = 4048 (Total 193 docs)
Test sentences = 3231 (Total 156 docs)

Our model used LinearSVC and the results are presented in table[2]. As we can see, we have achieved a F1 score of 86%.

Table 2. LinearSVC Scores

	Precision	Recall	F1Score
Dev Set	0.886	0.819	0.851
Test Set	0.891	0.833	0.861

4.2 LSTM-CRF

Model Description We have built a named entity recognizer which is often also considered as a sequence tagger. The model architecture involves Bi-LSTM and CRF. Additionally, it makes use of fasttext word embeddings for Spanish. It also builds word embeddings using character encodings. We have used word embeddings (fasttext) concatenated with model word embeddings (char based Bi-LSTM) while training the model. We then extract contextual representation of each word in a given sentence by running Bi-LSTM on the sentence. In the end, we used CRF to decode the output and get the category of each word.

Training Details In this section, we show a list of all hyper-parameters we used when training our model. This list includes optimizer, dropout, layer size, learning rate, learning rate decay, size of word embeddings, size of char embeddings among others. These parameters can be found in Table 3.

During the training of the model, we evaluated the model on the development set on each epoch and analyzed its performance through F1 score and confusion matrix generated. The number of epochs for which the model ran was decided by examining the F1 score of devset evaluations after every epoch and the training stopped when there was no improvement in the performance for consecutive epochs.

Table 3. Training Hyper Parameters

Hyper Parameter	**Value**
Optimizer	Adam
Dropout	0.5
Learning Rate	0.002
Learning Rate Decay	0.5
LSTM Hidden Size	200
Character Hidden Size	100
Word Embeddings	300
Character Embeddings	100

Results We ran the model with the above mentioned hyper-parameters and after running the model for 5 epochs, we obtained the confusion plot as shown in Fig 1.

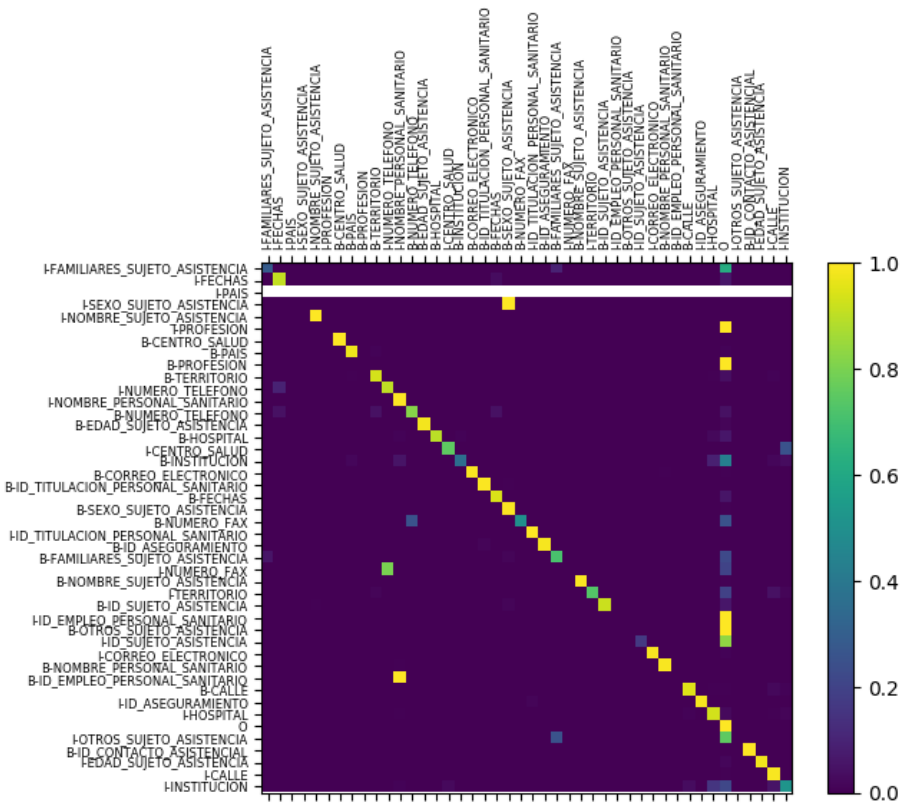

Fig. 1. Confusion Plot

Table 4 shows the precision, recall, and F1-scores we obtained on running the model on training, development and test data sets.

Table 4. LSTM-CRF Model Scores

	Precision	**Recall**	**F1-Score**
Train Set	0.985	0.939	0.962
Dev Set	0.960	0.911	0.935
Test set	0.957	0.913	0.935

5 Error Analysis

The neural model performed poorly for predicting the following types:

1. **Skewed distribution for category 'ID_SUJETO_ASISTENCIA':** As evident from confusion plot, we can see that the model is performing poorly

on the category 'ID_SUJETO_ASISTENCIA'. On analysis, we found out this category has a skewed distribution in the training set. In gold files, the category 'ID_SUJETO_ASISTENCIA' has been assigned to numbers in majority of the documents and only some times to text. Therefore, our model learnt to assign this category to numbers only.

2. **No heterogeneity in the data for category 'HOSPITAL':** On analysing the errors related to the category 'HOSPITAL', we found out that the label 'HOSPITAL' is assigned to those terms in the training data which contain the term 'HOSPITAL' in it. This means that we do not have variety of examples for this category in our training set. Thus, the model learnt to predict the category 'HOSPITAL' only if the token itself contains this term.

6 Conclusion

In this paper, we tried different approaches for the task of de-identification of PHI in Spanish clinical records. We started with building a LinearSVC model using various token-level features and static dictionaries of Spanish names and locations. We proposed a neural network that uses Bi-LSTM and CRF for named entity recognition. The neural model performed best for us on the given dataset.

7 Future Work

As a next step, we plan to build a system that is a combination of rule based model and our best performing neural model. Since the dataset consists of some structured text and some unstructured text in the medical documents, for the structured text, we will use the predictions made by a simple rule based system and for the unstructured text, we shall go with the neural model predictions.

References

1. Amund Tveit, Ole Edsberg, Thomas Brox Rst, Arild Faxvaag, ystein Nytr, Torbjrn Nordgrd, Martin Thorsen Ranang and Anders Grimsmo: Anonymization of General Practioner Medical Records. HelsIT, 5 pages (2019)
2. Gyrgy Szarvas, Richrd Farkas, Rbert Busa-fekete: State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. Journal of the American Medical Informatics Association Volume 14 Number 5, 7 pages (Sept / Oct 2007)
3. Kaung Khin, Philipp Burckhardt, Rema Padman: A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation. arXiv:1810.01570 [cs.CL], 15 pages (3 Oct 2018)
4. Xiao-Bai Li and Jialun Qin: Anonymizing and Sharing Medical Text Records. Inf Syst Res. Author manuscript, 47 pages (19 March 2018)
5. Yikun Guo, Robert Gaizauskas, Ian Roberts, George Demetriou, Mark Hepple: Identifying Personal Health Information Using Support Vector Machines. 5 pages (2006)

6. Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodrguez, Jose A Lopez Martin, Marta Villegas, Martin Krallinger: Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)
7. Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov: Learning Word Vectors for 157 Languages. Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)