Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled

Salvador Medina and Jordi Turmo
TALP Research Center - Universitat Politècnica de Catalunya
Carrer de Jordi Girona, 1-3, 08034 Barcelona
{smedina, turmo}@cs.upc.edu

Abstract

Electronic Health Records (EHR) are an important resource for the research and study of diseases, treatments and symptoms. However, due to data protection laws, information that could potentially compromise privacy must be anonymized before making use of them. Thus, the identification of these pieces of information is mandatory. This identification is usually performed by linguistic models built from EHRs corpora in which Protected Health Information (PHI) has been previously annotated. Nevertheless, two main drawbacks can occur. First, the annotated corpora required to build the models for a particular language may not exist. Second, unannotated corpora might exist for that language, containing very few words related to PHI mentions (i.e., very sparse population). In this situation, the process of manually annotating EHRs results extremely hard and costly, as PHI occurs in very few EHRs. This paper proposes an iterative method for building corpus with labelled PHI from a large unlabelled corpus with a very sparse population of target PHI. The method makes use of manually defined rules specified in the form of Augmented Transition Networks, and tries to minimize the seek of EHRs containing PHI, thus minimizing the cost of manually annotating very sparse EHRs corpora. We use the method with primary care EHRs written in Spanish and Catalan, although it is language-independent and could be applied to EHRs written in other languages. Direct and indirect evaluations performed to the resulting labelled corpus show the appropriateness of our method.

Keywords: sparse, anonymization, iterative method, Spanish, Catalan, health records

1. Introduction

The interest on identification Protected Health Information (PHI) in Electronic Health Records (EHR) with the objective of automatically de-identifying them has seen an important increment in recent years. For this reason, multiple PHI de-identification challenges have been issued, such as the Informatics for Integrating Biology to Bedside (i2b2) 2006 (Uzuner et al., 2007) and 2014 (Stubbs and Uzuner, 2015) or the 2016 CEGS N-GRID shared tasks (Stubbs et al., 2017). All of them focusing on the de-identification of English-written EHR following the guidelines by the Health Information Portability and Accountability Act (HIPAA).

The de-identification of PHI in health notes is indeed a very challenging task. To begin with, EHR are hard to parse, since they are often composed of unconnected observations in the form of short phrases containing severe syntactical and morphological errors. Moreover, PHI are usually uncommon, and can be easily confounded by procedures and drugs that are named after their developer. Because of this, general-purpose NER tools such as the dictionary-based *FreeLing* NER module are not appropriate for this task, and context-specific NER systems are required.

Thanks to the aforementioned challenges, multiple NER tools specifically crafted for PHI have been proposed. Supervised learning models such as the Bilinear Long Short-Time Memory (BiLSTM) network described by Dernoncourt (Dernoncourt et al., 2017) have managed to achieve remarkable results for PHI de-identification. However, with independency of the supervised model used, manually tagged corpora is mandatory for both training and evaluation. Moreover, this corpus should be big enough and representative of the diversity found in the unlabelled documents.

Several training corpora consisting of health records are available for English, most of them released for de-identification challenges, but others repurposed from de-identified health research datasets such as the MIMIC-II dataset (Saeed et al., 2002). Sadly, this is not the case for health records in languages other than English, for which labelled datasets are very limited. As a result, those state-of-the-art PHI de-identification supervised learning models cannot be adapted to health notes in other languages, due to the lack of annotated corpora.

Moreover, personal data protection laws, such as the Spanish *Ley de Protección de Datos*, do not allow researchers outside the health institutions to access identified health records. Consequently, the corpus must be manually labelled by the institution's personnel. However, due to the huge sparsity of PHI mentions in health notes (e.g. just less than 4 over 1000 words are names of person), the human annotators would be forced to check tens of thousands of documents to build a representative corpus. As a result, generating the needed training corpus can be prohibitively expensive for local health institutions.

In order to circumvent this drawback and make it cheaper for health institutions, we present an iterative method to build from scratch a corpus labelled with the occurrences of PHI. Inspired in active learning, the method selects relevant examples from unlabelled corpus using a set of manually defined search rules that is also enriched at each iteration. We used this method to build a bilingual Spanish/Catalan corpus with very sparse PHI labelled. Direct and indirect evaluations of the resulting labelled corpus are also reported.

The rest of the paper is structured as follows. Section 2. describes the iterative method used to select new relevant examples of PHI occurring in the unlabelled corpus. Sec-

tions 3. and 4. present the corpus to be labelled using our method, and both direct and indirect evaluations of the resulting labelled corpus, respectively. The results of these evaluations are described in Section 5.. Finally, Section 6. concludes.

2. The rule-based method

An strategy that is often used for building models when having a small labelled corpus but big sets of unlabelled data is active learning (Settles and Craven, 2008). Nevertheless, an initial training corpus is still required for building the initial supervised model, specially when applying state of the art models such as *Bilinear LSTM*, which may over-fit if the initial corpus is not large enough and active learning may not be successful as a result. Building such training corpus can be extremely time-consuming in the context of health records, since the density of some PHI categories such as names of people is significantly low (e.g., less than 0.28% of tokens in our corpus).

A possible alternative is to begin with a simpler manually defined rule-based system until the training corpus is big enough for a supervised model to be able to generalize from it. Regular expressions or gazetteer-based manual rules are commonly used for building simple initial models when not enough training data is available (Kozareva, 2006). The main issue with this approach, however, concerns diversity, as the examples obtained with a handcrafted rule-based system are usually similar among them and biased. This could be circumvented by defining several rules with minimal correlation, the main challenges being the ability to come up with diverse rules and knowing how many of them are needed.

Our approach revolves over the idea of starting from a diverse set of manually defined rules and defines an iterative methodology, inspired by active learning, for adding new rules to such set. New rules are defined and previous ones are refined with each iteration of our method so that the set keeps growing in complexity and diversity. We take profit of the expressiveness of Augmented Transition Networks (ATN) to be able to define and update such complex rules with ease.

2.1. The iterative method

The iterative method begins from an empty corpus. The first step is then to build a basic one by using text queries based on domain knowledge and gazetteers; and manually correcting a random sample of the retrieved documents, covering all possible categories. In our experiments, we began with 100 documents, but this would depend on the characteristics of the corpus and PHI categories.

With this basic corpus as a base, repeat the scheme below until the user is unable to come up with new rules given the requirements imposed to the F_1 score achieved by the rule set in the iterative training and validation corpora.

1. Run the set of rules against the training set and list the errors.
2. If a rule can be defined that covers more than one incorrect example in the training set without decreasing the F_1 score, add that rule. If no new rule can be defined, go to 4.
3. Evaluate using the validation set.
4. If recall is not increased and F_1 decreases, discard all the new rules and repeat from 2. If both precision and F_1 are decreased, update an existing rule so that precision increases in the training set and repeat from 3. Otherwise, repeat from 2.
5. Once no new rule can be added with the defined conditions, run the new set of rules against a subset of the unlabelled data.
6. Rank the new set of documents using the score function described in Section 2.2. and select those that are over the threshold. Repeat from 1.

2.2. Ranking and selection of new examples

The new set of documents that is added to the training set in each iteration is selected from the pool of documents by ranking them using the score function defined in equation 1 and discarding those below a given threshold score. We have designed this score function so that documents with multiple relevant instances are prioritized, specially those that belong to classes that are infrequent and hard to define manual rules for.

We determine the threshold score as the score that corresponds to the *elbow* point of the curve defined by the document's scores sorted in decreasing order. The *elbow criterion* is often used in cluster analysis to determine the optimal number of clusters so that adding another cluster does not give much better modeling of the data (Madhulatha, 2012). Similarly, in our case, we determine the number of selected documents so that including more does not add much more relevant examples.

$$f(d) = \sum_{i \epsilon K} N_i(d) * (1 - F_1(i)) * (1 - p_i)$$

$$p_i = \frac{\sum_{t \epsilon T} N_i(t)}{\sum_{i \epsilon K} \sum_{t \epsilon T} N_i(t)} \quad (1)$$

Where K is the set of classes, T is the set of documents in the training set, $N_i(d)$ is the number of examples of class i in document d and $F_1(i)$ is the F_1 score of class i in the validation set.

This score function prioritizes documents with multiple examples. What is more, the weight associated to each label decreases with increasing F_1 score for it - reaching 0 if $F_1 = 1.0$ - and a higher weight is assigned to labels that are uncommon in the training set.

2.3. The augmented transition networks

In our ATN implementation, sentences are parsed at token level by default, however we allow partial consumption of tokens via custom arc actions so that words with no spacing between them can be successfully identified.

Sentences are tokenized using the open-source *FreeLing 4.0* natural language processing suite (Padró and Stanilovsky, 2012), as it is the most feature-complete NLP

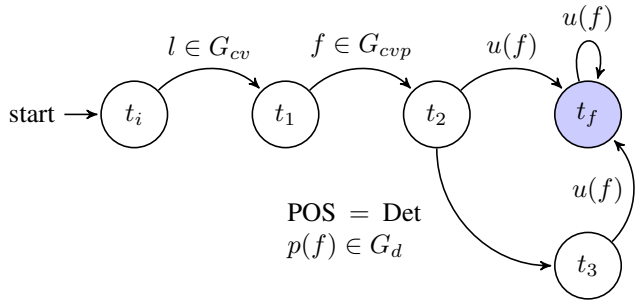


Figure 1: Example of an ATN rule. l, f and POS stand for lemma, form and Part of Speech respectively. $p(f)$ means to partially consume form f and $u(f)$ stands for uppercase. G_{cv}, G_{cvp} and G_d are gazetteers for communication verbs, communication verb pronouns and determinants. This rule can handle Examples **E.1** and **E.2**.

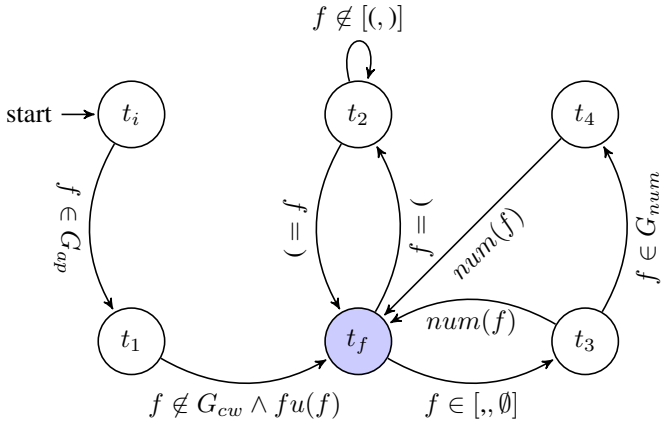


Figure 2: Fragment of an ATN rule. f stands for form. G_{ap} and G_{num} are gazetteers for addresses' prefixes and addresses' numbers prefixes respectively. $num(f)$ is a regular expression that determines whether f is a number or not. Maximum length of t_2 is limited to 5 tokens using edge actions and edge conditions. This rule can handle Examples **E.3** and **E.4**.

library for both Catalan and Spanish documents. However, due to the fact that *FreeLing 4.0* is optimized for texts written in standard language, the Named Entity Recognition and multi-word detection modules can lead to severe tokenization errors and we opted for disabling them.

Our ATNs can consume tokens based on their morphology, lemma or Part-of-Speech (POS) tag. Environmental variables can also be set based on the appearance of a certain token. In most of the cases, arcs check whether or not the token or sequence of tokens are included in a certain list of gazetteers, optionally adding restrictions relative to capitalization or POS tag.

The list below shows some examples of sentences that can be successfully parsed by some the of the rules that we have defined. A simplification of those rules are shown in Figures 1 and 2.

E.1 *Los derivo a bienestar social para hablar con Oliach.* (I derive them to social wellness to talk with Oliach).

E.2 *Parlo amb lAnna de la pauta a seguir.* (I talk to Anna about the guideline to follow).

E.3 *AVINGUDA MONTILIVI Nº 5 (al costat Suca-Mulla), tercer pis, porta D.* (5 MONTILIVI AVENUE (next to Suca-Mulla), third floor, door D.).

E.4 *AVENIDA DRASSENAS 17-21 TLF. 934416126.* (17-21 DRASSENAS AVENUE TEL. 934416126.). Note that *'DRASSENAS'* is misspelled, the correct form is *'Drassanes'*.

2.4. Observations

Given the characteristics of the iterative method that we propose, the resulting set of rules and labelled corpus ensures that:

- Rules that increase coverage (recall) are prioritized over those that increase precision, as the latter are not added unless the overall F_1 decreases.
- F_1 score increases monotonically in both training and validation set, first by increasing recall and then increasing precision in the latter iterations.
- New examples can be added indefinitely so that labelled sets of arbitrary size can be generated. As such, an additional stopping criteria should be defined. In this case, we stop once our set of rules surpasses 0.8 in the validation set.
- The resulting corpus does not maintain the proportions of entities found in the unlabelled corpus, since instances with a very low frequency and not easily identifiable are preferred.
- Documents with no entities are ignored so that the curators spend minimum time validating instances that do not contain relevant information.

The fact that the resulting corpus is unbalanced could potentially lead to worse performance in supervised learning models. Nevertheless, multiple methods have been presented over the last years that can be applied to unbalanced datasets so that this limitation is halved, which include semi-supervised algorithms (Huang and Kecman, 2004), re-sampling strategies (Liu et al., 2003) or weight adjusting mechanisms (Saerens et al., 2002).

3. The Spanish/Catalan corpus of health records

We apply the iterative method to a bilingual corpus of Electronic Health Records containing admission, progress, operative and discharge notes taken by doctors in the Catalan primary health care system. The goal is to be able to identify the Protected Health Information included in these records.

The *Institut Català de la Salut* (ICS) Primary Care Service's corpus of 2013 is composed by 12 files, each containing the short comments attached to the medical reports issued during the corresponding month of 2011. The notes are written in Spanish and Catalan, often combining words

	Full Corpus	Test	Validation
Notes	32882336	5000	311
Word occurrences	631020021	112281	15430
Tokens/Note	19.19	22.46	49.61
Words	3430167	33007	6346
Words ($F > 5$)	582047	2717	219
Spanish/Catalan	1.268:1	1.390:1	1.022:1

Table 1: Statistics of the Electronic Health Record corpus by the *Institut Català de la Salut* of 2013.

from both languages, and cover multiple fields: from common illnesses to psychology, dependency, drug use and so forth. Each record entry is identified by three numbers divided by vertical vars, but no additional structured information is provided. The first column in Table 1 summarizes some figures about the full unlabelled corpus. First row stands for the number of notes in the corpus; second and third rows refer to the number of word occurrences and the ratio of word ocurrences per note; fourth and fifth rows stand for the number of words without repetition with frequency greater than 1 and greater than 5, respectively; finally, the last row shows the ratio between Spanish and Catalan records within the corpus according to *FreeLing*'s language detection module.

3.1. Textual characteristics of the corpus

The documents in this corpus are written in natural language, usually composed of short sentences lacking verbal phrases and having severe non-gramatical morphological and syntactical phenomena. In addition to those, the list of phenomena listed below is recurrent in the corpus:

- Incoherent use of capitalization. For instance, *"realitzarem innmovilitzaació, recomanen e insisteim anar aH DE CALELLA PER CONFIRMAR FISURA I FRACTURA, DIU QUE NO HI ANIRÀ QUE NO VOL ESPERAR-SE 4 H.P:Realitzem inmovilització i control en una seetmana."* combines fully lowercased phases with fully uppercased ones.

- Use of contractions. An example of this can be found in the sentence *"Pac que finaliza tto"*, where the words *Pac* and *tto* are used instead of *Paciente* (patient) and *tratamiento* (treatment).

- Use of punctuation marks instead of spaces or lack of them. For example, in the sentence *"Algun subcrepitante en bases...Normas.Pulmicort-100 2-1(15 dias)."*, the words *bases*, *Normas* and *Pulmicort-100* are not spaced. What is more, in sentence *"Controlada HVhebron anualment."*, *HVhebron* should be *H. V. Hebron*, as it refers to *Hospital Vall Hebron*.

- Enumerations of measures and readings from medical analysis. For example, *"Usa L/C OD 85º-0.50 +1.00 0.8 /+4.00. OI 115º-1.00 +0.25 0.9 /+3.50.AO 4DP BT en VL.Rx ¿OD NG. OI NG Ad/3.00."*

- Inconsistent use of languages, since notes often combine Spanish and Catalan words, phrases or idioms. For instance, sencence *"M:febre de 39ºC tot el dia a pesar que la mare li ha donat Dalsy, vomits i mucositat nasal."* is written in Catalan but includes the Spanish expression *a pesar que* (despite of), while sentence *"E:herida mordida palma de mano D.P:neteja, steristrip..."* is written in Spanish but uses the Catalan verb *neteja* (to clean).

3.2. Protected Health Information categories

The PHI categories that we consider in this work follow the de-identification directives given by the *Institut Universitari d'Investigació en Atenció Primària* (IDIAP), a medical research center subordinated to the *Institut Català de la Salut* (ICS). A health note is considered successfully de-identified if the PHI entities listed below are replaced by their respective category name. Estimations of the proportions of tokens corresponding to each PHI category are given based on the observation of a subset of documents.

1. PERSON: Name or surname of a patient, relative, medical staff or any other person mentioned in the report. (about 0.28% of the tokens).

2. LOCATION: Physical locations or geographic subdivisions including street address, city, county, precinct, ZIP code, et cetera. This also includes public locations such as hospitals, clinics, schools and others. (about 1.12% of the tokens, 0.73% being public locations).

3. TELEPHONE: Digits of a phone number. (below 0.01% of the tokens).

4. EMAIL: E-mail address. (less than 0.01%).

5. DNI: Spanish *Documento Nacional de Identificación*. (less than 0.01% of the tokens).

6. SOCIAL_SECURITY_ID: Spanish social security number. (less than 0.01% of the tokens).

7. SANITARY_CARD_ID: Catalan sanitary card number. (less than 0.01% of the tokens).

It is also worth noting that there is a high degree of correlation between PHI in the health records. Based on the observation of a subset of documents, more than 50% of health records containing PHI include multiple instances, 44% of them having 3 or more. While the probability that a note contains any PHI is below 9%. As could be expected, health records that explicitly include personal information of a patient or doctor such as the name often include other information such as the names of relatives and partners, as well as working places, clinics etcetera.

4. Evaluation Framework

In order to evaluate the method that we are presenting, we apply both direct and indirect evaluation. First, we evaluate how the manual validation time by curators is optimized in terms of the faction of relevant examples presented to them. Additionally, we indirectly evaluate the corpus that is obtained during the iterative method in terms of the quality of a model trained with it.

	Validation	Test	Resulting Corpus
PERSON	372	282	699
LOCATION	99	680	825
TELEPHONE	7	6	17
Notes	311	5000	1051
Notes /w PHI	299	667	793

Table 2: Count of instances of PHI corresponding to categories PERSON, LOCATION and TELEPHONE in corpora

We restrict evaluation to the identification of instances of PHI that correspond to categories PERSON and LOCATION, even though the iterative process is applied to every category described in Section 3.2.. Categories TELEPHONE, EMAIL, DNI, SOCIAL_SECURITY_ID and SANITARY_CARD_ID have a formal structure and previous work in the subject has proven that simple regular expressions are enough to cover all instances (Yang and Garibaldi, 2015). Moreover, as shown in Section 3.2., the density of such entities in the full corpus reported by IDIAP is so low that instances in the evaluation corpus may not be representative enough. For these two reasons, we have opted for neglecting them in the evaluation figures.

4.1. Validation and testing partitions

Our method starts from a completely empty labelled corpus which grows at each iteration, jointly with the set of rules. In order to be able to evaluate each modification the set of rules, we have previously selected and manually labelled a small validation set of documents from the unlabelled corpus. Considering that our main goal is to embrace as much diversity as possible and we want to keep evaluation time as small as possible, this validation set is composed of just positive examples. These examples are selected by skimming the set of unlabelled documents and selecting those in which an example is spotted in order to lower the required building time.

The test set, which is used to perform the indirect evaluation of the final set of rules obtained after the iterative method, is composed of 5000 randomly selected documents from the whole set of unlabelled ones. Opposite to the resulting labelled corpus and the validation set, the test set maintains the proportion and density of PHI mentions of the unlabelled corpus.

Columns 2 and 3 of Table 1 show statistics about the number of health records and words in the Validation and Test corpora. Columns 1 and 2 of Table 2 list the amount of instances of each category of PHI that we evaluate in our work for these two corpora.

4.2. Direct evaluation of the guided labelling process

A way to measure the fraction of PHI in the health notes that are manually labelled while ensuring that an heterogeneous set of examples are retrieved is to look at the F_1 score. On the one hand, we would like the number of documents that are supposed to contain PHI to actually contain them, which means that the generated rules should be precise. On the other hand, relevant examples should not be ignored, so the rules are required to have a high *recall*. F_1 computes the harmonic mean of both these scores so it is the most suitable evaluation measure for this criteria. Note that strict evaluation must not be enforced, since the health notes are supposed to be manually labelled anyway. Exact matching of the entities' bounds is not mandatory and they will be considered as *true positive* if the labels contain any of the entities' tokens.

We evaluate the initial and final rule sets obtained after applying the iterative method using a test corpus composed of 5000 randomly selected and manually labeled health records described in Section 4.1. according to the evaluation criteria described above. Additionally, we compare the F_1 score achieved by the aforementioned context-specific sets of rules to the one using the general-purpose *Named Entity Recognition and Classification* (NERC) module included in *FreeLing*.

4.3. Indirect evaluation of the resulting corpus

The indirect evaluation of the corpus obtained as a result of the iterative process is done by using it as the training set of supervised PHI identification models. We train a Conditional Random Field (CRF) sequence tagger using morphological, part-of-speech, lemmas and clustered word-embedding input features, which is a widely-used model for PHI identification capable of achieving state of the art performance in recent de-identification shared tasks (Yang and Garibaldi, 2015), (Dehghan et al., 2015).

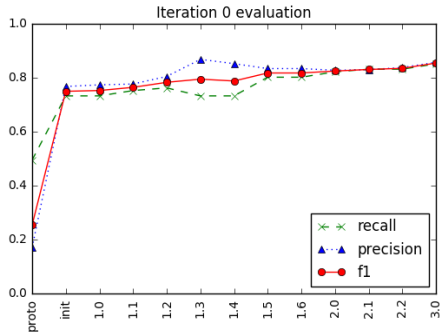
We compare the models trained with the iterative corpus to others trained using a larger corpus generated by randomly selecting and labelling examples from the unlabelled corpus. In particular, we take the test corpus disposed for the direct evaluation of the rule set and divide it into 8 folds, which are then re-purposed as 8 test corpora and 8 training corpora of 625 and 4375 health records respectively. The iteratively generated corpus is evaluated with each one of these test folds independently.

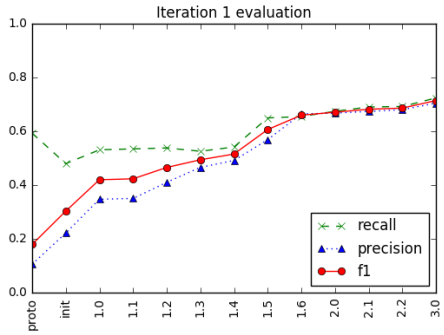
.

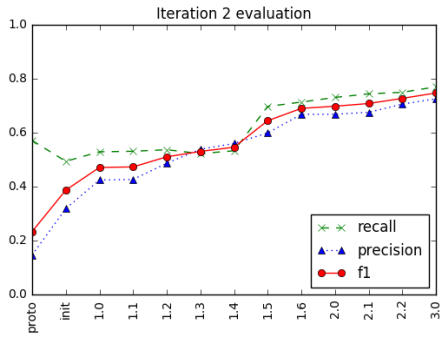
5. Results

Figure 3 shows the evolution of *recall*, *precision* and F_1 score evaluated using the training corpora obtained after iterations 0 to 2 for each update of the rule set. Given the conditions imposed by the iterative process, F_1 score in the validation set increases monotonically. Both *recall* and *precision* also have an ascending trend. This means that the set of rules is improved at each iteration, being able to cover a broader variety of common contexts of PHI while avoiding mislabelled instances.

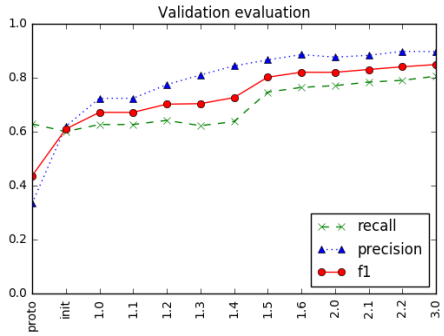
Table 3 shows the direct evaluation of *recall*, *precision* and F_1 scores in the test corpus for the *FreeLing NERC* module, as well as for the initial and final sets of rules. F_1 score is considerably lower than in the validation set, due to the fact that the latter only includes health records containing instances of PHI whereas the test corpus maintains the proportions of the full unlabelled corpus. The final recall is over 70% while precision is around 50%. This means that the set of rules is capable of retrieving training corpora including 70% of the instances of PHI in the unlabeled corpus showing 50% of false positives. Hence it can







(a) Iterative training corpora



(b) Validation corpus

Figure 3: Evaluation results for classes PERSON, LOCATION and overall for the iterative corpora built for iterations 0 to 2 (Figure 3a) and the validation corpus (Figure 3b).

considerably reduce the time required for the manual labelling process while discarding just 30% of the positive instances. Even though the iterative method achieves similar F_1 score for categories PERSON and LOCATION (0.564 and 0.509 respectively), the behaviour of *recall* and *precision* differs considerably. *Recall* for category PERSON is high compared to LOCATION (0.772 and 0.371 respectively), probably due to the fact that rules for LOCATION are less abundant but more precise, since they rely more in gazetteers.

	Eval.	NERC	initial	final
ALL	Recall	0.052	0.147	0.702
	Prec.	0.494	0.208	0.489
	F_1	0.094	0.172	0.576
PERSON	Recall	0.436	0.676	0.772
	Prec.	0.023	0.196	0.445
	F_1	0.044	0.304	0.564
LOCATION	Recall	0.517	0.013	0.371
	Prec.	0.064	0.127	0.809
	F_1	0.114	0.024	0.509

Table 3: Evaluation results in the test set for the general-purpose *Freeling* NERC module, and for the initial and final sets of hand-crafted rules.

	Eval.	Cross-Val.	Res. Corpus
ALL	Recall	0.721 (0.027)	0.699 (0.042)
	Prec.	0.839 (0.026)	0.769 (0.047)
	F_1	0.774 (0.017)	0.732 (0.039)
PERSON	Recall	0.784 (0.064)	0.759 (0.093)
	Prec.	0.909 (0.041)	0.730 (0.061)
	F_1	0.840 (0.025)	0.744 (0.057)
LOCATION	Recall	0.695 (0.040)	0.676 (0.056)
	Prec.	0.812 (0.022)	0.783 (0.061)
	F_1	0.748 (0.037)	0.726 (0.052)

Table 4: Mean *recall*, *precision* and F_1 score obtained by a CRF model trained using the labelled corpus obtained after 3 iterations of the method (1051 health records) compared to the *8-fold* cross validation of the test corpus (4350 health records) for the 8 testing partitions. Standard deviation is shown between brackets.

Table 4 shows the mean *recall*, *precision* and F_1 score of the indirect evaluation of the resulting labelled corpus after 3 iterations, compared to the 8-fold cross-validation of the test corpus used for direct evaluation. F_1 score using the iterative corpus is a 0.042 points lower compared to the traditional corpus, achieving similar *recall* (0.022 points lower) but significantly worse *precision* (0.07 points lower). This remarkable downgrade in *precision* is expectable, as the corpus has a higher density of positive examples. Nevertheless, the obtained results are promising, since they show that it is possible achieve similar *recall* after just 3 iterations. This leads us to believe that with more iterations and unsupervised re-sampling strategies to increase *precision*, the iteratively generated corpus could outperform the traditional one.

6. Conclusions

In this paper, we describe our method to build a corpus, in which protected information is labelled, from a large set of unlabelled electronic health records containing very sparse relevant information. Basically, hand-crafted rules are iteratively created for the automatic labelling of new protected information occurring in the health records. The re-

trieved documents that are considered most informative are selected and manually corrected in order to be used later to design new hand-crafted rules and refine the existing ones. Using this method, we created a bilingual Spanish/Catalan health records corpus with labelled protected information. We evaluated the resulting corpus in two ways: a direct evaluation by examining the relevance of the rules created at each iteration, and an indirect evaluation by comparing CRFs models learned using the resulting labelled corpus and a manually labeled extract of the full unlabelled corpus.

Given that we get comparable results using the iteratively generated corpus while requiring much less manual effort in terms of documents to be validated, we believe that the proposed method is appropriate for building inexpensive PHI identification training corpora. This more efficient use of resources is specially significant in subsequent iterations, as the proportion of PHI in the retrieved health notes increases and fetch rules grow in complexity.

We conclude that the presented method is a reasonable alternative to the much expensive process of uninformedly labelling random health records to build corpora when the density of target entities is low, while making minimal compromises to the final performance of the supervised models trained with it.

7. Acknowledgments

This works has been partially funded by the Spanish Goverment and by the European Union through GRAPHMED project (TIN2016-77820-C3-3-R and AEI/FEDER,UE.)

References

Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., and Nenadic, G. (2015). Combining knowledge-and data-driven methods for de-identification of clinical narratives. *Journal of biomedical informatics*, 58:S53–S59.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Huang, T. M. and Kecman, V. (2004). Semi-supervised learning from unbalanced labeled data–an improvement. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 802–808. Springer.

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the eleventh conference of the European chapter of the association for computational linguistics: student research workshop*, pages 15–21. Association for Computational Linguistics.

Liu, B., Dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building text classifiers using positive and unlabeled examples. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 179–186. IEEE.

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Saeed, M., Lieu, C., Raber, G., and Mark, R. G. (2002). Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics.

Stubbs, A. and Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Stubbs, A., Filannino, M., and Uzuner, Ö. (2017). De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Yang, H. and Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of biomedical informatics*, 58:S30–S38.