UNIVERSITY OF TWENTE

MASTER THESIS

Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records

Author
Jan Trienes

Supervisors
Dr. Djoerd Hiemstra
Dr. Christin Seifert

External Supervisor
Dr. Dolf Trieschnigg (Nedap)

June 2019

Faculty of Electrical Engineering, Mathematics and Computer Science
Department of Computer Science, Specialization Data Science

Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records

Jan Trienes[1], **Dolf Trieschnigg**[2], **Christin Seifert**[1], **Djoerd Hiemstra**[1]
[1] University of Twente, Enschede, Netherlands
[2] Nedap Healthcare, Groenlo, Netherlands

Abstract

Unstructured information in electronic health records provide an invaluable resource for medical research. To protect the confidentiality of patients and to conform to privacy regulations, de-identification methods automatically remove personally identifying information from these medical records. However, due to the unavailability of labeled data, most existing research is constrained to English medical text and little is known about the generalizability of de-identification methods across languages and domains. In this study, we construct a novel dataset consisting of the medical records of 1260 patients among three domains of Dutch healthcare. We test the generalizability across languages and domains for three de-identification methods. Our experiments show that an existing rule-based method specifically developed for the Dutch language fails to generalize to this new data, and that a state-of-the-art neural architecture outperforms rule-based and feature-based methods when testing on new domains even when limited training data is available.

1 Introduction

With the strong adoption of electronic health records (EHR), large quantities of unstructured medical patient data become available. This data offers significant opportunities to advance medical research and to improve healthcare related services. However, it has to be ensured that the privacy of a patient is protected when performing secondary analysis of medical data. This is not only an ethical prerequisite, but also a legal requirement imposed by privacy legislations such as the US Health Insurance Portability and Accountability Act (HIPAA) and the European General Data Protection Regulation (GDPR). To facilitate privacy protection, de-identification has been proposed as a process that removes or masks any kind of protected health information (PHI) of a patient such that it becomes difficult to establish a link between an individual and the data (Meystre, 2015). What type of information constitutes PHI is in part defined by privacy laws of the corresponding country. For instance, the HIPAA regulation defines 18 categories of PHI including names, geographic locations, and phone numbers (HIPAA, 2012). According to the HIPAA safe-harbor rule, data is no longer personally identifying and subject to the privacy regulation if these 18 PHI categories have been removed.

As most EHRs consist of unstructured, free-form text, manual de-identification is a time-consuming and error-prone process (Neamatullah et al., 2008; Douglass et al., 2004) which does not scale to the amounts of data needed for many data mining and machine learning scenarios. Therefore, automatic de-identification methods are desirable. Previous research proposed a wide range of methods that make use of natural language processing techniques including rule-based matching and machine learning (Meystre, 2015). However, most evaluations are constrained to medical records written in the English language. The generalizability of de-identification methods across languages and domains is largely unexplored.

To test the generalizability of existing de-identification methods, we annotated a new dataset of 1260 medical records from three sectors of Dutch healthcare: elderly care, mental care and disabled care (Section 3). We then compare the performance of the following three de-identification methods on this data (Section 4):

1. A rule-based system named DEDUCE developed for Dutch psychiatric clinical notes (Menger et al., 2018)
2. A feature-based Conditional Random Field (CRF) as described in Liu et al. (2015)
3. A deep neural network with a bidirectional

long short-term memory architecture and a CRF prediction layer (BiLSTM-CRF) (Akbik et al., 2018)

We test the transferability of each method across three domains of Dutch healthcare. Finally, the generalizability of the methods is compared across languages using two widely used English benchmark corpora (Section 5).

This paper makes two main contributions. First, our experiments show that an existing rule-based de-identification method for the Dutch language fails to generalize to a new dataset of Dutch medical records. Second, we offer a novel comparison of several state-of-the-art de-identification methods both across languages and domains. Our experiments show that a popular neural architecture generalizes best even when limited amounts of training data are available. The neural method only considers word/character sequences which we find to be sufficient and more robust across languages and domains compared to the structural features employed by traditional machine learning approaches. However, our experiments also reveal that the neural method may still experience a substantially lower performance in new domains. A direct consequence for de-identification practitioners is that pre-trained models require additional fine-tuning to be fully applicable to new domains.

The implementation of the de-identification systems, pre-trained models and code for running the experiments is available at: `github.com/nedap/deidentify`

2 Related Work

Previous work on de-identification can be roughly organized into four groups: (1) creation of benchmark corpora, (2) approaches to de-identification, (3) work on languages other than English and (4) cross-domain de-identification.

Various English benchmark corpora have been created including nursing notes, longitudinal patient records and psychiatric intake notes (Neamatullah et al., 2008; Stubbs and Uzuner, 2015; Stubbs et al., 2017). Furthermore, Deléger et al. (2012) created a heterogeneous dataset of documents of 22 different types. While most existing datasets only contain records of a single medical institute, our dataset includes records from 9 different institutions that are active in the Dutch healthcare sector. Similar to the corpus created by Deléger et al. (2012), our dataset consists of heterogeneous documents that significantly vary in structure, content and writing style.

Most existing de-identification approaches are either rule-based or machine learning based. Rule-based methods combine various heuristics in form of patterns, lookup lists and fuzzy string matching to identify PHI (Gupta et al., 2004; Neamatullah et al., 2008). The majority of machine learning approaches employ feature-based CRFs (Aberdeen et al., 2010; He et al., 2015), ensembles combining CRFs with rules (Stubbs et al., 2015a) and most recently also neural networks (Dernoncourt et al., 2017; Liu et al., 2017). A thorough overview of the different de-identification methods is given in Meystre (2015). In this study, we compare several state-of-the-art de-identification methods. With respect to rule-based approaches, we apply DEDUCE, a recently developed method for Dutch data (Menger et al., 2018). For a feature-based machine learning method, we re-implement the token-level CRF by Liu et al. (2015). Previous work on neural de-identification used a BiLSTM-CRF architecture with character-level and ELMo embeddings (Dernoncourt et al., 2017; Khin et al., 2018). Similarly, we use a BiLSTM-CRF but apply recent advances in neural sequence modeling by using contextual string embeddings (Akbik et al., 2018).

To the best of our knowledge, we are the first study to offer a comparison of de-identification methods across languages. With respect to de-identification in other languages, only two studies consider Dutch data. Scheurwegs et al. (2013) apply a Support Vector Machine and a Random Forest classifier to a dataset of 200 clinical records. Menger et al. (2018) developed and released a rule-based method on 400 psychiatric nursing notes and treatment plans of a single Dutch hospital. Furthermore, de-identification in several other languages has been studied including German, French, Korean and Swedish (Richter-Pechanski et al., 2018; Névéol et al., 2018).

With respect to cross-domain de-identification, the 2016 CEGS N-GRID shared task evaluated the portability of pre-trained de-identification methods to a new set of English psychiatric records (Stubbs et al., 2017). Overall, the existing systems did not perform well on the new data. Here, we provide a similar comparison by cross-testing on three domains of Dutch healthcare.

Datset	Our Corpus	i2b2	Nursing
Language	Dutch	English	English
Documents	1260	1304	2434
Patients	1260	296	148
Tokens	448,795	1,057,302	444,484
Vocabulary	25,429	36,743	19,482
PHI categories	16	32	10
PHI instances	17,464	28,872	1779
Median PHI/doc.	9	18	0

Table 1: Overview of the datasets used in this study.

3 Datasets

This section describes the construction of our Dutch benchmark corpus. The data was sampled from 9 healthcare institutes and annotated for PHI according to a tagging scheme derived from Stubbs and Uzuner (2015). Furthermore, following common practice in the preparation of de-identification corpora, we replaced PHI instances with realistic surrogates to comply with privacy regulations. To compare the performance of the de-identification methods across languages, we use the English i2b2/UTHealth and the nursing notes corpus (Stubbs and Uzuner, 2015; Douglass et al., 2004). An overview of the three datasets can be found in Table 1.

3.1 Data Sampling

We sample data from a snapshot of the databases of 9 healthcare institutes with a total of 83,000 patients. Three domains of healthcare are equally represented in this snapshot: elderly care, mental care and disabled care. We consider two classes of documents to sample from: surveys and progress reports. Surveys are questionnaire-like forms which are used by the medical staff to take notes during intake interviews, record the outcomes of medical tests or to formalize the treatment plan of a patient. Progress reports are short documents describing the current conditions of a patient receiving care, sometimes on a daily basis. The use of surveys and progress reports differs strongly across healthcare institute and domain. In total, this snapshot consists of 630,000 surveys and 13 million progress reports.

When sampling from the snapshot described above, we aim to maximize both the variety of document types, and the variety of PHI, two essential properties of a de-identification benchmark corpus (Deléger et al., 2012). First, to ensure a wide variety of document types, we select surveys in a stratified fashion according to their type label provided by the EHR system (e.g., intake interview, care plan, etc.). Second, to maximize the variety in PHI, we sample medical reports on a patient basis: for each patient, a random selection of 10 medical reports is combined into a patient file. We then select patient files uniformly at random which ensures that no patient appears multiple times within the sample. Furthermore, to control the annotation effort, we impose two subjective limits on the document length. A document has to contain at least 50 tokens, but no more than 1000 tokens to be included in the sample. For each of the 9 healthcare institutes, we sample 140 documents (70 surveys and 70 patient files), which yields a total sample size of 1260 documents (see Table 1).

Category	i2b2 (Stubbs and Uzuner, 2015)	Our Tags
Name	Patient, Doctor, Username	Name Initials
Profession	Profession	Profession
Location	Room, Department	Internal Location
	Hospital, Organization	Hospital, Organization Care Institute
	Street, City, State, ZIP, Country	Address
Age	Over 90, Under 90	Age
Date	Date	Date
Contact	Phone, FAX, Email	Phone/FAX, Email
	URL, IP	URL/IP
IDs	SSN, 8 fine-grained ID tags	SSN, ID
Other	Other	Other

Table 2: PHI tags used to annotate our dataset. The tagging scheme was derived from the i2b2 tags.

We received approval for the collection and use of our dataset from the ethics review board of our institution.

3.2 Annotation Scheme

Since the GDPR does not provide any strict rules about which types of PHI should be removed during de-identification, we base our PHI tagging scheme on the guidelines defined by the US HIPAA regulations. In particular, we closely follow the annotation guidelines and the tagging scheme used by Stubbs and Uzuner (2015) which consists of 32 PHI tags among 8 classes: *Name*, *Profession*, *Location*, *Age*, *Date*, *Contact Information*, *IDs* and *Other*. The *Other* category is used for information that can be used to identify a patient, but it does not fall into any of the remaining categories. For example: the sentence *"the patient was a guest speaker on diabetes in the Channel 2 talkshow."* would be tagged as *Other*. It is worth mentioning that this tagging scheme does not only capture direct identifiers relating to a patient (e.g., name and date of birth), but also indirect identifiers that could be used in combination with other information to reveal the identity of a patient. Indirect identifiers include, for example, the doctor's name, information about the hospital and a patient's profession.

We made two adjustments to the tagging scheme by Stubbs and Uzuner (2015). First, to reduce the annotation effort, we merged some of the 32 fine-grained PHI tags to a more generic set of 16 tags (see Table 2). For example, the fine-grained location tags *Street*, *City*, *State*, *ZIP*, and *Country* were merged into a generic *Address* tag. While this simplifies the annotation process, it complicates the generation of realistic surrogates. Given an address string, one has to infer its format to replace the individual parts with surrogates of the same semantic type. We address this issue in Section 3.4. Second, due to the high frequency of care institutes in our dataset, we decided to introduce a separate *Care Institute* tag that complements the *Organization* tag. This allows for a straightforward surrogate generation where names of care institute are replaced with another care institute rather than with more generic company names (e.g., Google).

3.3 Annotation Process

Following previous work on the construction of de-identification benchmark corpora (Stubbs and Uzuner, 2015; Deléger et al., 2012), we employ a double-annotation strategy: two annotators read and tag the same documents. In total, 12 non-domain experts annotated the sample of 1260 medical records independently and in parallel. The documents were randomly split into 6 sets and we randomly assigned a pair of annotators to each set. To ensure that the annotators had a common understanding of the annotation instructions, an evaluation session was held after each pair of annotators completed the first 20 documents. The annotators took in total 77h to double-annotate the entire corpus of 1260 documents, or approximately 3.6 minutes per document. We measured the inter-annotator-agreement (IAA) using entity-level F1 scores[1]. Table 3 shows the IAA per PHI category. Overall, the agreement level is fairly high. However, we find that location names (i.e., care institutes, hospitals, organizations and internal locations) are often highly ambiguous which is reflected by the low agreement scores of these categories.

PHI Tag	Count	Frac. (%)	IAA
Name	9558	54.73	0.96
Date	3676	21.05	0.86
Care Institute	997	5.71	0.52
Initials	778	4.45	0.46
Address	748	4.28	0.75
Organization	712	4.08	0.38
Internal Location	242	1.39	0.29
Age	175	1.00	0.39
Profession	122	0.70	0.31
ID	114	0.65	0.43
Phone/Fax	97	0.56	0.93
Email	95	0.54	0.94
Hospital	92	0.53	0.42
Other	33	0.19	0.03
URL/IP	23	0.13	0.70
SSN	2	0.01	0.50
Total	17,464	100	0.84

Table 3: Distribution of PHI tags in our corpus. The inter-annotator agreement (IAA) as measured by the micro-averaged F1 score is shown per category.

To improve annotation efficiency, we integrated the rule-based de-identification tool DEDUCE (Menger et al., 2018) with our annotation software to pre-annotate each document. This functionality could be activated on a per-document basis by each annotator. If an annotator used this functionality, they had to review the pre-annotations, correct potential errors and check for missed PHI instances. During the evaluation sessions, annotators mentioned that the existing tool proved helpful when annotating repetitive names, dates and email addresses. Note that this pre-annotation strategy might give DEDUCE a slight advantage. However, the low performance of DEDUCE in the formal benchmark in Section 5 does not reflect this.

After annotation, the main author of this paper reviewed 19,165 annotations and resolved any disagreements between the two annotators to form the gold-standard of 17,464 PHI annotations. Table 3 shows the distribution of PHI tags after adjudication. Overall the adjudication has been done risk-averse: if only one annotator identified a piece of text as PHI, we assume that the other annotator has missed this potential PHI instance. In addition to the manual adjudication, we performed two automatic checks: (1) we ensured that PHI instances

---

[1] It has been shown that the F-score is more suitable to quantify IAA in sequence-tagging scenarios compared to other measures such as the Kappa score (Deléger et al., 2012).

occurring in multiple files received the same PHI tag, and (2) any instances that were tagged in one part of the corpora but not in the other were manually reviewed and added to the gold-standard. We used the BRAT annotation tool for both annotation and adjudication (Stenetorp et al., 2012). We include our annotation guidelines and example PHI annotations in Appendix E.

3.4 Surrogate Generation

As the annotated corpus consists of personally identifying information which is protected by the GDPR, we generate artificial replacements for each of the PHI instances before using the data for the development of de-identification methods. This process is known as surrogate generation, a common practice in the preparation of de-identification corpora (Stubbs et al., 2015b). As surrogate generation will inevitably alter the semantics of the corpus to an extent where it affects the de-identification performance, it is important that this step is done as thoroughly as possible (Yeniterzi et al., 2010). Here, we follow the semi-automatic surrogate generation procedure that has been used to prepare the i2b2/UTHealth shared task corpora. Below, we summarize this procedure and mention the language specific resources we used. We refer the reader to Stubbs et al. (2015b) for a thorough discussion of the method. After running the automatic replacement scripts, we reviewed each of the surrogates to ensure that continuity within a document is preserved and no PHI is leaked into the new corpus.

A list of 10,000 most common family names and given names is used to generate random surrogates for name PHI instances[2]. We replace dates by first parsing the format (e.g., "12 nov. 2018" → "%d %b. %Y")[3], and then randomly shifting all dates within a document by the same amount of years and days into the future. For addresses, we match names of cities, streets, and countries with a dictionary of Dutch locations[4], and then pick random replacements from that dictionary. As Dutch ZIP codes follow a standard format ("1234AB"), their replacement is straightforward. Names of hospitals, care institutes, organizations and internal locations are randomly shuffled within the corpus. PHI instances of type *Age* are capped at 89 years. Finally, alphanumeric strings such as *Phone/FAX*, *Email*, *URL/IP*, *SSN* and *IDs* are replaced by substituting each alphanumeric character with another character of the same class. We manually rewrite *Profession* and *Other* tags, as an automatic replacement is not applicable.

4 Methods

This section presents the three de-identification methods and the evaluation procedure.

4.1 Rule-based Method: DEDUCE

DEDUCE is an unsupervised de-identification method specifically developed for Dutch medical records (Menger et al., 2018). It is based on lookup tables, decision rules and fuzzy string matching and has been validated on a corpus of 400 psychiatric nursing notes and treatment plans of a single hospital. Following the authors' recommendations, we customize the method to include a list of 1200 institutions that are common in our domain. Also, we resolve two incompatibilities between the PHI coding schemes of our dataset and the DEDUCE output. First, as DEDUCE does not distinguish between hospitals, care institutes, organizations and internal locations, we group these four PHI tags under a single *Named Location* tag. Second, our *Name* annotations do not include titles (e.g., "Dr.", "Ms."). Therefore, titles are stripped from the DEDUCE output.

4.2 Feature-based Method: Conditional Random Field

CRFs and hybrid rule-based systems provide state-of-the-art performance in recent shared tasks (Stubbs et al., 2015a, 2017). Therefore, we implement a CRF approach to contrast with the unsupervised rule-based system. In particular, we re-implement the token-based CRF method by Liu et al. (2015) an re-use a subset[5] of their features (see Table 4). The linear-chain CRF is trained using LBFGS and elastic net regularization (Zou and Hastie, 2005). Using a validation set, we optimize the two regularization coefficients of the L_1 and L_2 norms with a random search in the log_{10} space of $[10^{-4}, 10^1]$. We use the *CRFSuite* implementation by Okazaki (2007).

4.3 Neural Method: BiLSTM-CRF

To reduce the need for hand-crafted features in traditional CRF-based de-identification, recent work

---

[2] See `www.naamkunde.net`, accessed 2019-05-15

[3] Rule-based date parser: `github.com/jeffreystarr/dateinfer`, accessed 2019-05-15

[4] See `openov.nl`, accessed 2019-05-15

[5] We disregard word-representation features as Liu et al. (2015) found that they had a negative performance impact.

Group	Description
Bag-of-words (BOW)	Token unigrams, bigrams and trigrams within a window of $[-2, 2]$ of the current token.
Part-of-speech (POS)	Same as above but with POS n-grams.
BOW + POS	Combinations of the previous, current and next token and their POS tags.
Sentence	Length in tokens, presence of end-mark such as '.', '?', '!' and whether sentence contains unmatched brackets.
Affixes	Prefix and suffix of length 1 to 5.
Orthographic	Binary indicators about word shape: is all caps, is capitalized, capital letters inside, contains digit, contains punctuation, consists of only ASCII characters.
Word Shapes	The abstract shape of a token. For example, "7534-Df" becomes "####-Aa".
Named-entity recognition	NER tag assigned by the spaCy tagger.

Table 4: Features used by the CRF method. The features are identical to the one by Liu et al. (2015), but we exclude word-representation features.

applies neural methods (Liu et al., 2017; Dernoncourt et al., 2017; Khin et al., 2018). Here, we re-implement a BiLSTM-CRF architecture with contextual string embeddings, which has recently shown to provide state-of-the-art results for sequence labeling tasks (Akbik et al., 2018). Hyperparameters are set to the best performing configuration in Akbik et al. (2018): we use stochastic gradient descent with no momentum and an initial learning rate of 0.1. If the training loss does not decrease for 3 consecutive epochs, the learning rate is halved. Training is stopped if the learning rate falls below 10^{-4} or 150 epochs are reached. Furthermore, the number of hidden layers in the LSTM is set to 1 with 256 recurrent units. We employ locked dropout with a value of 0.5 and use a mini-batch size of 32. With respect to the embedding layer, we use the pre-trained GloVe (English) and fasttext (Dutch) embedding on a word-level, and concatenate them with the pre-trained contextualized string embeddings included in Flair[6] (Pennington et al., 2014; Grave et al., 2018; Akbik et al., 2019).

4.4 Preprocessing

We use a common preprocessing routine for all three datasets. For tokenization and sentence segmentation, the spaCy tokenizer is used[7]. The POS/NER features of the CRF method are generated by the built-in spaCy models. After sentence segmentation, we tag each token according to the Beginning, Inside, Outside (BIO) scheme. In rare occasions, sequence labeling methods may produce invalid transitions (e.g., `O-` → `I-`). In a post-processing step, we replace invalid `I-` tags with `B-` tags (Reimers and Gurevych, 2017).

4.5 Evaluation

The de-identification methods are assessed according to precision, recall and F1 computed on an entity-level, the standard evaluation approach for NER systems (Tjong Kim Sang and De Meulder, 2003). In an entity-level evaluation, predicted PHI offsets and types have to match exactly. Following the evaluation of de-identification shared tasks, we use the micro-averaged entity-level F1 score as primary metric (Stubbs et al., 2015a)[8].

We split our corpus and the nursing notes corpus into train/validation/test sets with a 60/20/20 ratio. As the i2b2 corpus has a pre-defined test set of 40%, a random set of 20% of the training documents serves as validation data. Finally, we test for statistical significance using two-sided approximate randomization (Yeh, 2000).

5 Results

This section presents our experimental results. An overview can be found in Table 5.

5.1 De-identification of Dutch Dataset

Both machine learning methods outperform the rule-based system DEDUCE by a large margin (see Table 5). Furthermore, the BiLSTM-CRF provides a substantial improvement of 10% points in recall over the traditional CRF method, while maintaining precision. Overall, the neural method has an entity-level recall of 87.1% while achieving a recall of 95.6% for names, showing that the neural method is operational for many de-identification scenarios. In addition, we make the following observations.

Neural method performs at least as good as rule-based method. By inspecting the model performance on a PHI-tag level, we observe that

---

[6] `github.com/zalandoresearch/flair`, accessed 2019-05-15

[7] `spacy.io`, accessed 2019-05-15

[8] De-identification systems are often also evaluated on a less strict token-level. As a system that scores high on an entity-level will also score high on a token-level, we only measure according to the stricter level of evaluation.

	Our Corpus (Dutch)			i2b2 (English)			Nursing Notes (English)		
Method	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
DEDUCE	0.807	0.564	0.664	-	-	-	-	-	-
CRF	**0.919**$^{\blacktriangle}$	0.775$^{\blacktriangle}$	0.841$^{\blacktriangle}$	0.952	0.796	0.867	**0.914**	0.685	0.783
BiLSTM-CRF	0.917°	**0.871**$^{\blacktriangle}$	**0.893**$^{\blacktriangle}$	**0.959**$^{\blacktriangle}$	**0.869**$^{\blacktriangle}$	**0.912**$^{\blacktriangle}$	0.886°	**0.797**$^{\blacktriangle}$	**0.839**$^{\blacktriangle}$

Table 5: Evaluation summary: micro-averaged scores are shown for each dataset and method. Statistically significant improvements over the score on the previous line are marked with $^{\blacktriangle}$($p < 0.01$), and $^{\circ}$ depicts no significance. The rule-based method DEDUCE is not applicable to the English datasets.

	BiLSTM-CRF		DEDUCE	
PHI Tag	Prec.	Rec.	Prec.	Rec.
Name	**0.965**	**0.956**	0.849	0.805
Date	**0.926**	**0.920**	0.857	0.441
Initials	**0.828**	**0.624**	0.000	0.000
Address	**0.835**	**0.846**	0.804	0.526
Age	**0.789**	**0.732**	0.088	0.122
Profession	**0.917**	**0.262**	0.000	0.000
ID	**0.800**	**0.480**	0.000	0.000
Phone/Fax	0.889	**1.000**	**0.929**	0.812
Email	0.909	**1.000**	**1.000**	0.900
Other	0.000	0.000	0.000	0.000
URL/IP	**1.000**	**0.750**	0.750	**0.750**
Named Loc.	**0.797**	**0.659**	0.279	0.058
Care Institute	0.686	0.657	n/a	n/a
Organization	0.780	0.522	n/a	n/a
Internal Loc.	0.737	0.509	n/a	n/a
Hospital	0.778	0.700	n/a	n/a

Table 6: Entity-level precision and recall per PHI category on our Dutch corpus. Scores are compared between the rule-based tagger DEDUCE (Menger et al., 2018) and the BiLSTM-CRF model. The *Named Loc.* tag is the union of the 4 specific location tags which are not supported by DEDUCE. Tags are ordered by frequency with location tags fixated at the bottom.

the neural method outperforms DEDUCE for all classes of PHI (see Table 6). Only for the *Phone* and *Email* category, the rule-based method has a slightly higher precision. This suggests that in most environments where training data are available (or can be obtained), the machine learning methods are to be preferred.

Rule-based method can provide a "safety net." It can be observed that DEDUCE performs reasonably well for names, phone numbers, email addresses and URLs (see Table 6). As these PHI instances are likely to directly reveal the identity of an individual, their removal is essential. However, DEDUCE does not generalize beyond the PHI types mentioned above. Especially named locations are non-trivial to capture with a rule-based system as their identification strongly relies on the availability of exhaustive lookup lists. In contrast, the neural method provides a significant improvement for named locations (5.8% vs. 65.9% recall). We assume that word-level and character-level embeddings provide an effective tool to capture these entities.

Initials, IDs, professions are hard to detect. During annotation, we observed a low F1 annotator agreement of 0.46, 0.43, and 0.31 for initials, IDs and professions, respectively. This shows that these PHI types are among the hardest to identify, even for humans (see Table 3). One possible cause for this is that IDs and initials are often hard to discriminate from abbreviations and medical measurements. We observe that the BiLSTM-CRF detects those PHI classes with high precision but low recall. With respect to professions, we find that phrases are often wrongly tagged. For example, colloquial job descriptions (e.g., "works behind the cash desk") as opposed to the job title (e.g., "cashier") make it infeasible to tackle this problem with lookup lists, while a machine learner likely requires more training data to capture this PHI.

5.2 Cross-language De-identification

When training and testing both machine learning methods on the English i2b2 and the nursing notes datasets, we can observe that the BiLSTM-CRF significantly outperforms the CRF in both cases (see Table 5). Similar to our Dutch dataset, the neural method provides an increase of up to 11.2% points in recall (nursing notes) while the precision remains relatively stable. This shows that the neural method has the best generalization capabilities even across languages. More importantly, it does not require the development of domain-specific lookup lists or sophisticated pattern matching rules. To put the results into perspective: the second-

	Training Domain		
Method	Elderly	Disabled	Mental
DEDUCE	0.683	0.565	0.675
CRF	0.414	0.697	0.719
BiLSTM-CRF	**0.775**	**0.775**	**0.839**

Table 7: Summary of the transfer learning experiment on our Dutch dataset. Each method is trained on data of one care domain and tested on the other two domains. All scores are micro-averaged entity-level F1.

	Training Domain		
Test Domain	Elderly	Disabled	Mental
Elderly	0.746	0.698	0.703
Disabled	0.796	0.919	0.879
Mental	0.744	0.806	0.871

Table 8: Detailed performance analysis of the BiLSTM-CRF method in the transfer learning experiment. In-domain test scores are shown on the diagonal. All scores are micro-averaged entity-level F1.

highest ranked team in the i2b2 2014 challenge used a sophisticated ensemble combining a CRF with domain-specific rules (Stubbs et al., 2015a). Their system obtained an entity-level F1 score of 0.9124 which is on-par with the performance of our neural method that requires no configuration. We can expect that the performance of the neural method further improves after hyperparameter optimization. Finally, note that both machine learning methods can easily be applied to a new PHI tagging scheme, whereas rule-based methods are limited to the PHI definition they were developed for.

5.3 Cross-domain De-identification

In many de-identification scenarios, heterogeneous training data from multiple medical institutes and domains are rarely available. This raises the questions, how well a model that has been trained on a homogeneous set of medical records generalizes to records of other medical domains. We trained the three de-identification methods on one domain of Dutch healthcare (e.g., elderly care) and tested each model on the records of the remaining two domains (e.g., disabled care and mental care). Table 7 summarizes the performance of each method on the different tasks. We follow the same training and evaluation procedures described in Section 4.5.

Again, the neural method consistently outperforms the rule-based and feature-based methods in all three domains which suggests that it is a fair default choice for de-identification. This is underlined by the fact that the amount of training data is severely limited in this experiment: each domain only has 420 documents of which 20% of the records are reserved for testing. Interestingly, DEDUCE performs rather stable and even outperforms the CRF within the domain of elderly care.

Given an ideal de-identification method, one would expect that performance on unseen data of a different domain is similar to the test score obtained on the available (homogeneous) data. Table 8 shows a performance breakdown for each of the three testing domains for the neural method. It can be seen that in 4 out of 6 cases, the test score in a new domain is lower than the test score obtained on the in-domain data. The largest delta of the observed in-domain test score (disabled care, 0.919 F1) and the performance in the transfer domain (elderly care, 0.698 F1) is 0.221 in F1 which raises an important point when performing de-identification in practice. While the neural method overall provides the best generalization capabilities compared to the other de-identification methods, the performance can still be significantly lower when applying a pre-trained model in new domains.

6 Conclusion

This paper presents the construction of a novel Dutch dataset and a comparison of state-of-the-art de-identification methods across Dutch and English medical records. Our experimental results show the following: (1) An existing rule-based method for the Dutch language does not generalize well to the new data. (2) If one is looking for an out-of-the-box de-identification method, neural approaches show the best generalization performance across languages and domains. (3) When testing across different domains, a substantial decrease of performance has to be expected, an important consideration when applying de-identification in practice.

There are several directions for future work. Motivated by the limited generalizability of pre-trained models across different domains, transfer learning techniques can provide a way forward. A preliminary study by Lee et al. (2018) shows that they can be beneficial for de-identification. Finally, our experiments show that phrases such as professions are among the most difficult information to de-identify. It is an open challenge how to design methods that can capture this type of information.

References

John S. Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Benjamin Wellner, Cheryl Clark, David A. Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE identification scrubber toolkit: Design, training, and assessment. *I. J. Medical Informatics*, 79(12):849–859.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Louise Deléger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnár, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, Illinois, USA, November 3-7, 2012*.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *JAMIA*, 24(3):596–606.

Margaret Douglass, Gari D Clifford, Andrew Reisner, George B Moody, and Roger G Mark. 2004. Computer-assisted de-identification of free text in the MIMIC II database. In *Computers in Cardiology, 2004*, pages 341–344. IEEE.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Dilip Gupta, Melissa Saul, and John Gilbertson. 2004. Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. *American Journal of Clinical Pathology*, 121(2):176–186.

Bin He, Yi Guan, Jianyi Cheng, Keting Cen, and Wenlan Hua. 2015. CRFs based de-identification of medical records. *J. of Biomedical Informatics*, 58(S):S39–S46.

HIPAA. 2012. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. *US Department of Health and Human Services, Washington, DC) Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html*. Accessed on 2019-05-15.

Kaung Khin, Philipp Burckhardt, and Rema Padman. 2018. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *Computing Research Repository*, arXiv:1810.01570.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 4470–4473, Miyazaki, Japan.

Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, and Suisong Zhu. 2015. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics*, 58:S47–S52.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *J. of Biomedical Informatics*, 75(S):S34–S42.

Vincent Menger, Floor Scheepers, Lisette Maria van Wijk, and Marco Spruit. 2018. DEDUCE: A pattern matching method for automatic de-identification of dutch medical text. *Telematics and Informatics*, 35(4):727 – 736.

Stephane M. Meystre. 2015. De-identification of unstructured clinical data for patient privacy protection. In Aris Gkoulalas-Divanis and Grigorios Loukides, editors, *Medical Data Privacy Handbook*, pages 697–716. Springer International Publishing.

Ishna Neamatullah, Margaret M. Douglass, Li-Wei H. Lehman, Andrew T. Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC Med. Inf. & Decision Making*, 8:32.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *J. Biomedical Semantics*, 9(1):12:1–12:13.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. *Computing Research Repository*, arXiv:1707.06799.

Phillip Richter-Pechanski, Stefan Riezler, and Christoph Dieterich. 2018. De-identification of german medical admission notes. *Studies in health technology and informatics*, 253:165–169.

Elyne Scheurwegs, Kim Luyckx, Filip Van der Schueren, and Tim Van den Bulcke. 2013. De-identification of clinical free text in dutch with limited training data: A case study. In *Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013*, pages 18–23.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1. *Journal of Biomedical Informatics*, 75:S4 – S18.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015a. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits. 2015b. Challenges in synthesizing surrogate PHI in narrative EMRs. In Aris Gkoulalas-Divanis and Grigorios Loukides, editors, *Medical Data Privacy Handbook*, pages 717–735. Springer International Publishing.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reyyan Yeniterzi, John S. Aberdeen, Samuel Bayer, Benjamin Wellner, Lynette Hirschman, and Bradley Malin. 2010. Effects of personal identifier resynthesis on clinical text de-identification. *JAMIA*, 17(2):159–168.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

A Supplemental Material

This appendix provides additional analyses and background information for the paper. It includes an error analysis of the best performing model (Section B) and a practical experiment illustrating the precision/recall tradeoff (Section C). Furthermore, we discuss how automatic de-identification can be beneficial for Nedap, a Dutch company that develops software solutions for the healthcare market (Section D). Finally, we include the annotation guidelines that were used to annotate our de-identification benchmark corpus (Section E).

B Error Analysis

To gain a better understanding of the best performing model and an intuition for its limitations, we conduct a manual error analysis of the false positives (FPs) and false negatives (FNs) produced by the BiLSTM-CRF. This analysis allows to identify possible error sources and avenues for future work. We discuss the error categorization scheme in Section B.1 and present the results in Section B.2.

B.1 Error Categorization

We distinguish between two main error groups: (1) modeling errors and (2) annotation/preprocessing errors. We define modeling errors to be problems that can be addressed with different de-identification techniques and additional training data. In contrast, annotation and preprocessing errors are not directly caused by the sequence labeling model, but are issues in the training data or the preprocessing pipeline which need to be addressed manually. Inspired by the classification scheme of Dernoncourt et al. (2017), we consider the following sources of modeling errors:

- **Abbreviation:** PHI instances which are abbreviation/acronyms for names, care institutes and companies. These are hard to detect and can be ambiguous as they are easily confused with medical terms and measurements.
- **Ambiguity:** A human reader may be unable to decide whether the text fragment is PHI.
- **Debatable:** It can be argued that the token should not have been annotated as PHI.
- **Prefix:** Names of internal locations, organizations and companies are often prefixed with articles (i.e., "de" and "het"). Sometimes, it is unclear whether the prefix is part of the official name or part of the sentence construction. This ambiguity is reflected in the training data which causes the model to inconsistently include/exclude those prefixes.
- **Common Language:** PHI instances which consist of common-language. These are hard to discriminate from the surrounding text.
- **Other:** Remaining modeling errors that do not fall into the categories mentioned above. In those cases, it is not immediately apparent why the misclassification occurs.

Preprocessing errors are categorized as follows:

- **Missing Annotation:** The text fragment should have been annotated as PHI, but was missed during the annotation phase.
- **Annotation Error:** The annotator assigned an invalid entity boundary.
- **Tokenization Error:** The annotated text span could not be split into a compatible token span. Those tokens were marked as "Outside (O)" during BIO tagging.

All error categories are mutually exclusive.

B.2 Results of Error Analysis

Table 9 summarizes the error analysis results and shows the absolute and relative frequency of each error category. Overall, we find that the majority of modeling errors cannot be easily explained through human inspection ("Other reason" in Table 9). The remaining errors are mainly caused by ambiguous PHI instances and preprocessing errors. In more detail, we make the following observations:

Abbreviations are the second most common cause for modeling errors (13.9% of FNs, 9.7% of FPs). We hypothesize that more training data will likely not in itself help to correctly identify this type of PHI. It is conceivable to design custom features (e.g., based on shape, positioning in a sentence, presence/absence in a medical dictionary) to increase precision. However, it is an open question how recall can be improved.

PHI instances consisting of common language are likely to be wrongly tagged (7.5% FNs, 3.1% FPs). This is caused by the fact that there are insufficient training examples where common language is used to refer to PHI. For example, the organization name in the sentence "Vandaag Beter Horen gebeld" (Eng: "Called Hear Better today") was incorrectly classified as non-PHI. Each individual word, and also the combination of the two words,

	FNs ($n = 469$)		FPs ($n = 288$)	
Category	Count	Part	Count	Part
Model Errors				
Abbreviation	65	13.9%	28	9.7%
Ambiguity	15	3.2%	7	2.4%
Debatable	7	1.5%	4	1.4%
Prefix	10	2.1%	10	3.5%
Common language	35	7.5%	9	3.1%
Other reason	275	58.6%	159	55.2%
Annotation/Preprocessing Errors				
Missing Annotation	-	-	33	11.5%
Annotation Error	21	4.5%	18	6.3%
Tokenization Error	41	8.7%	20	6.9%
Total	469	100%	288	100%

Table 9: Summary of the manual error analysis of false negatives (FNs) and false positives (FPs) produced by the BiLSTM-CRF. All error categories are mutually exclusive.

can be used in different contexts without referring to PHI. However, in this specific context, it is apparent that “Beter Horen” must refer to an organization.

A substantial amount of errors is due to annotation and preprocessing issues. Annotation errors (4.5% FNs, 6.3% FPs) can be resolved by correcting the respective PHI offsets in the gold standard. Tokenization errors (8.7% FNs, 6.9% FPs) need to be fixed through a different preprocessing routine. For example, the annotation `<DATE 2016>/<DATE 2017>` should have been split into `[2016, /, 2017]` with BIO tagging `[B, O, B]`. However, spaCy segmented this text into a single token `[2016/2017]`. In this case, entity boundaries do no longer align with token boundaries which results in an invalid BIO tagging of `[O]` for the entire span.

Several false positives are in fact PHI and should be annotated. The model identifies several PHI instances which were missed during the annotation phase (11.5% of the FPs). Once more, this demonstrates that proper de-identification is an error-prone task for human annotators.

B.3 Training Examples vs. Performance

To gain a better understanding of the model performance, we study the impact of the training dataset size on the detection quality of each PHI tag. Figure 1 visualizes the number of training examples in relation to the F1 score of the BiLSTM-CRF for each PHI category. Generally, and as expected, it can be observed that a large number of training examples correlates with a high F1 score (e.g., *Name* and *Date* tags). However, few training examples do not necessarily correspond to a low F1 score. In line with the findings in Dernoncourt et al. (2017), PHI types which vary significantly in shape and content (e.g., *Organization*) require substantially more training examples to be correctly identified than PHI types with little variation (e.g., *Email*). For instance, the *Phone* and *Email* PHI tags have relatively few training examples (less than 100), but an F1 score that is just as high as the F1 score for *Name* (more than 6000 examples).

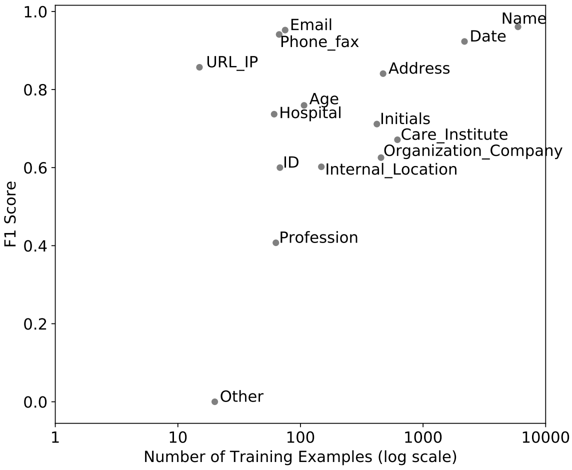


Figure 1: Impact of the number of training examples on the F1 score of the BiLSTM-CRF. Generally, more training examples contribute to higher F1 scores.

B.4 Example Misclassifications

For illustration purposes, Table 10 provides example PHI instances that were missed by the BiLSTM-CRF. There are two interesting observations. First, the neural method may fail to identify PHI in a new and unseen context. For example, the date “woensdag” (Eng: “Wednesday”) has been successfully de-identified in other contexts. Furthermore, we can find several instances where common language causes misclassification (see examples for Organization, Age and in Table 10).

C Precision/Recall Tradeoff

So far, our model analysis considered precision and recall of a de-identification method to be equally important. However, this is rarely the case when applying de-identification in practice where a downstream task determines the correct tradeoff between the two measures. For example, when sharing med-

PHI Category	FNs	Support	Examples
Name	85	1941	Wel wou bas dat vader aan tafel kwam zitten.
Date	64	803	Tevens mut einde zorg per 310167 voor bg-ind. Notitie gestuurd deelgenomen aan de verlate kerstlunch, heeft hiervan genoten. Afspraak is vastgelegd voor volgende week woensdag
Care Institute	74	216	We hebben samen gekeken of de ZORGCO hem nog op een ander Afspraak met woonzorgnet is niet goed gegaan. Vandaag bij de RAAK geweest. Was ok.
Initials	67	178	N.a.v. overleg KB met trajectbegeleiders is besloten dat 10/02/70: GE: @ KB naar EoR: Vanmiddag telefonisch gesproken
Address	24	156	contact gehad met Florencia van der Voort. (WMO Made)
Organization	65	136	Samen CJIB gebeld om navraag te doen over de OV-boetes. Vandaag Beter Horen gebeld inverband met
Internal Location	27	55	Wonend in de twee vieren van het Huize Padua.
Profession	31	42	Heeft fietsenmaker en Dienstmeisje opleidingen afgerond. Sinds juni 2069 ander werk (inkomensconsulent), heeft geen last
Age	11	41	een man van (bijna) halverwege de tachtig. Hij heeft de eerste 45 jaar van zijn leven in relatieve anarchie
ID	13	25	20/08/71 KG: MUT ZZP AAT BG-ind Kl 5 voor 88100691 en PV toewijzen naar De Kameleon op AGB 78449083.
Phone/Fax	0	16	*No misclassified samples*
Email	0	10	*No misclassified samples*
Hospital	3	10	Overdracht mst:geen cognitieve problemen
URL/IP	1	4	Met vriendelijke groet, https://www.gymmas.nl/tzepcwtnmy Greta
Other	4	4	De televisie uitzendingen brengen veel teweeg bij haar

Table 10: Example entities (underlined) which were missed by the BiLSTM-CRF. Number of false negatives (FNs) and number of test instances (Support) is shown for each PHI category.

ical records with an external research institute, no PHI should be leaked. Thus, high levels of recall are desirable even if it is at the expense of precision.

Below, we illustrate how a pre-trained model with a probabilistic output can be operationalized by introducing a classification threshold T that acts as a parameter to control the precision/recall trade-off. First, we consider the binary PHI prediction task: given a token x, decide whether or not it is PHI. Let $P(Y = PHI|x)$ depict the probability of token x to be PHI. By default, we obtain the outcome y as follows:

$$y = \begin{cases} \text{PHI}, & \text{if } P(Y = PHI|x) \geq T \\ \text{non-PHI}, & \text{otherwise}, \end{cases}$$

where $T = 0.5$. In the multiclass setting, the probability of $P(Y = PHI|x)$ is the sum of the probabilities for all PHI classes. In case of the BiLSTM-CRF model, a softmax function is used to obtain the probability distribution over all PHI tags from the CRF transition matrix. If we shift the classification threshold T, precision/recall for the positive class (i.e., PHI) can be adjusted: a higher value of T corresponds to an increase in precision, and a lower value of T to an increase in recall.

C.1 Thresholding in Binary PHI Detection

Figure 2 shows the precision/recall curves for the binary PHI prediction problem over all possible classification thresholds. Two operating points are highlighted: a recall level of 0.95 and a recall level of 0.99. Observe that precision remains relatively

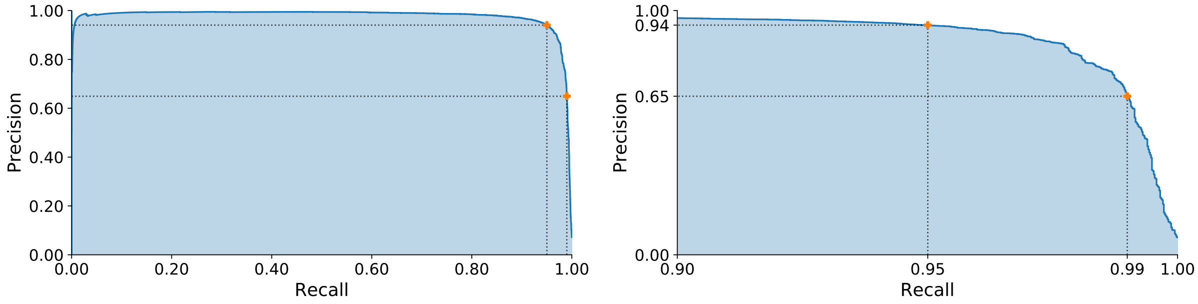


Figure 2: Token-level precision vs. recall over all possible classification thresholds for the binary PHI detection problem. Scores are for the positive (PHI) class and are reported for the test set. Two operating points are highlighted: the classification thresholds which satisfy a recall level of 0.95 and 0.99, respectively. The right figure is zoomed to the recall range $[0.9, 1]$.

	Recall			Precision		
PHI Category	$T = .5$	$T(\text{Rec.} = .95)$	$T(\text{Rec.} = .99)$	$T = .5$	$T(\text{Rec.} = .95)$	$T(\text{Rec.} = .99)$
Address	0.89	0.89 ±0.00	0.90 +0.01	0.90	0.90 −0.01	0.76 −0.14
Age	0.82	0.87 +0.05	0.92 +0.10	0.70	0.65 −0.05	0.25 −0.45
Care Institute	0.72	0.77 +0.04	0.86 +0.14	0.74	0.73 −0.01	0.48 −0.26
Date	0.98	0.98 ±0.00	0.99 +0.01	0.97	0.96 −0.01	0.84 −0.13
Email	1.00	1.00 ±0.00	1.00 ±0.00	0.91	0.83 −0.08	0.77 −0.14
Hospital	0.79	0.86 +0.07	0.93 +0.14	0.50	0.46 −0.04	0.18 −0.32
ID	0.56	0.68 +0.12	0.84 +0.28	0.70	0.65 −0.05	0.12 −0.58
Initials	0.65	0.69 +0.04	0.91 +0.26	0.83	0.81 −0.02	0.41 −0.42
Internal Loc.	0.57	0.59 +0.02	0.68 +0.10	0.75	0.70 −0.06	0.26 −0.50
Name	0.97	0.97 ±0.00	0.99 +0.02	0.97	0.96 ±0.00	0.92 −0.05
Organization	0.55	0.61 +0.06	0.65 +0.10	0.71	0.67 −0.04	0.30 −0.41
Other	0.00	0.00 ±0.00	0.00 ±0.00	0.00	0.00 ±0.00	0.00 ±0.00
Phone/Fax	1.00	1.00 ±0.00	1.00 ±0.00	0.97	0.95 −0.02	0.79 −0.18
Profession	0.58	0.68 +0.10	0.90 +0.32	0.85	0.75 −0.10	0.12 −0.73
SSN	0.00	0.00 ±0.00	0.00 ±0.00	0.00	0.00 ±0.00	0.00 ±0.00
URL/IP	0.75	0.75 ±0.00	1.00 +0.25	1.00	1.00 ±0.00	0.80 −0.20
Non-PHI	0.997	0.996 −0.001	0.968 −0.029	0.996	0.997 +0.001	0.999 +0.003
PHI (Binary)	0.935	**0.950** +0.015	**0.990** +0.055	0.954	0.941 −0.014	0.649 −0.305

Table 11: Token-level precision and recall of the BiLSTM-CRF at three operating points. $T(\text{Rec.} = .95)$ and $T(\text{Rec.} = .99)$ denote the classification thresholds at which a recall of 0.95 and 0.99 is obtained on the test set in the binary PHI prediction problem. For reference, the gray numbers indicate the performance delta compared to the default classification threshold of $T = 0.5$.

stable approximately until a recall level of about 0.95. Starting from there, precision experiences a sharp drop. At a 0.99 recall level, precision dropped significantly to 0.65. Note that there is no consensus about what level of recall is acceptable for de-identification. However, 0.95 has been frequently quoted (Stubbs and Uzuner, 2015) which is why we are highlighting this operating point along with an even stricter level of recall.

C.2 Effects of Thresholding on PHI-tag Level

One might ask, to what extend is the detection quality of each PHI tag affected? Table 11 illustrates precision and recall on a PHI tag level for three settings of T: the default threshold $T = 0.5$, and the thresholds at which a recall of 0.95 and 0.99 is obtained on the test set. In this case, we generalize the binary-PHI prediction as follows. If the sum of the probabilities for all PHI-tags for a given token, $P(Y = PHI|x)$, is greater than the threshold T, we assign the most likely PHI tag to that token.

The largest changes in performance can be observed for the PHI classes where the model originally performed poorly (e.g., Initials). On the contrary, PHI classes for which the classifier already performed good (e.g., Names), do not experience substantial changes when adjusting the threshold. Furthermore, we see that the recall of the non-PHI class (i.e., tokens tagged as "Outside" in the BIO scheme) drops from 0.997 to 0.968 when applying the strictest threshold with respect to the PHI detection. At a first glance, this change of 3 percent points seems rather negligible. However, note that potentially important medical information might be removed during this step. Depending on the downstream task, one needs to evaluate the quality of the remaining text and whether it is still suitable for the task at hand.

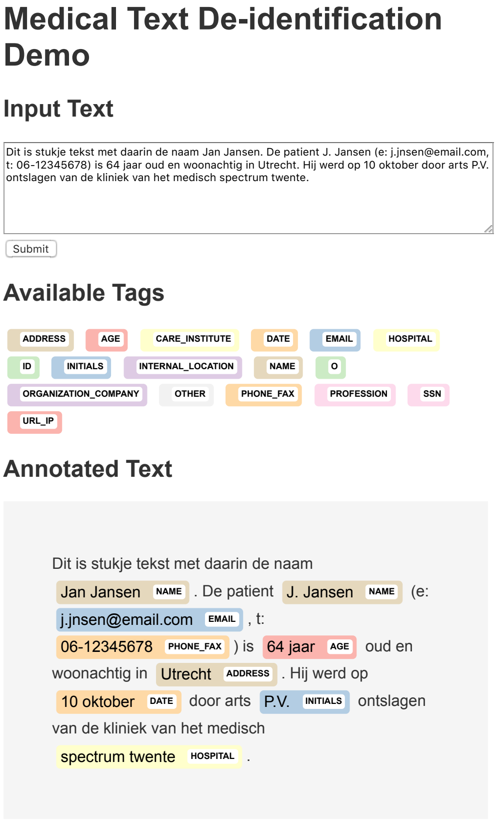
D De-identification at Nedap

This section briefly discusses the relevance of de-identification techniques for Nedap Healthcare, a company that provides software solutions to the Dutch healthcare sector.

D.1 Use Cases

Nedap envisions to use the de-identification methods developed in this paper in three broader areas:

- **Machine learning and data analysis.** Privacy protection is a major concern when building statistical models in the context of data mining and machine learning. Prior de-identification of the input data reduces the risk that privacy sensitive information is leaked into the statistical models. To give an example, the removal of PHI can help to avoid biases in a language generating model. Similarly, it can help to ensure that PHI (e.g., a name) does not become a discriminating feature in a document classification task.
- **Creation of realistic training and development software environments.** A major factor that limits effective user training and user experience design is the lack of realistic software environments. Currently, development environments contain nonsensical and artificial data which has limited utility. Existing snapshots of a customer database cannot be used for training and development purposes as they contain confidential medical information about the clients. However, when combining anonymization techniques for structured data (e.g., k-anonymity) with the methods developed in this paper, realistically looking replicas of existing applications can be created while drastically reducing the privacy impact.



Dit is stukje tekst met daarin de naam Jan Jansen. De patient J. Jansen (e: j.jnsen@email.com, t: 06-12345678) is 64 jaar oud en woonachtig in Utrecht. Hij werd op 10 oktober door arts P.V. ontslagen van de kliniek van het medisch spectrum twente.

Dit is stukje tekst met daarin de naam Jan Jansen NAME . De patient J. Jansen NAME (e: j.jnsen@email.com EMAIL , t: 06-12345678 PHONE_FAX) is 64 jaar AGE oud en woonachtig in Utrecht ADDRESS . Hij werd op 10 oktober DATE door arts P.V. INITIALS ontslagen van de kliniek van het medisch spectrum twente HOSPITAL .

Figure 3: Screenshot of our tool that we developed to demonstrate the de-identification method.

- **Connecting with academia and research institutes.** We aim to make the de-identification software alongside with the trained models publicly available. This will help Nedap Healthcare to connect with the community of de-identification practitioners, research communities and offers the potential for future collaborations on this subject.

Besides that, this work contributes towards the privacy protection practices that are implemented within Nedap and will help to fulfill the obligations set out by the GDPR.

D.2 Demo

We developed a demo application to allow users to experiment with the de-identification pipeline. It consists of two parts: (1) a Python web service that exposes both the preprocessing pipeline and the de-identification method as a REST API, and (2) a lightweight JavaScript client to interact with this API and to visualize sequence tagging results (see Figure 3).

D.3 Future Work at Nedap Healthcare

There are two major directions for future work when it comes to applying de-identification in practice at Nedap Healthcare.

1. **Combine neural and rule-based method.** The error analysis in Section B shows that the neural method offers little explainability and may fail to recognize identical PHI occurring in a different context. For that reason, an ensemble combining a rule-based method with the neural method may offer additional assurances that are useful in practice. For example, a lookup list can be used to ensure that if a PHI instance occurs which matches an entry in the lookup list exactly, it will be removed regardless of the context.
2. **Assess impact on downstream tasks.** Second, the impact of de-identification on a number of downstream tasks of Nedap Healthcare would make an interesting experiment. In particular, one should assess how much downstream tasks are affected under different operating points for the precision/recall tradeoff (see Section C). This can lead to general recommendations about which operating points are suitable for a specific application.

E Annotation Guidelines

For the development of an automatic de-identification software, we require medical records where the protected health information (PHI) has been marked up so that the annotations can be used to develop automatic de-identification methods. The annotated data is the type information that must be removed/replaced from a patient record in order to be considered de-identified. We defined 8 categories of PHI that can relate to a patient, but also to relatives, employers, household members or the doctor of a patient. In total 16 tags can be assigned to a piece of text:

1. NAME
    - Name
    - Initials
2. PROFESSION (not of medical staff)
3. LOCATION
    - Hospital
    - Care Institute (Zorgorganisatie)
    - Organization/Company
    - Address
    - Internal location (e.g., building code, room, floor)
4. AGE
5. DATE
6. CONTACT
    - Phone/FAX
    - Email
    - URL/IP-address
7. ID
    - Social security number (SSN/BSN)
    - Any other ID number
8. OTHER

E.1 Overall Annotation Rules

When annotating, the following rules apply:

1. When tagging something that is PHI but it is not obvious what to tag it as, think about what it should be replaced with and whether that will make sense in the document (“replacement test”).
2. When in doubt whether something is *tag A* or *tag B*, annotate it as the most likely tag and add a note to the annotation.
3. When in doubt, annotate! We do not want to miss PHI.

To give an example of the replacement test, consider this sentence:

Example E.1. *In 2015 is hij met de andere cliënten verhuisd naar de woonvoorziening Kerklaan in Bennebroek.*

[Translated] *In 2015 he moved with other clients to the housing facility Kerklaan in Bennebroek.*

It is clear that the housing facility has been named after its location “Kerklaan.” So instead of annotating “Kerklaan” as an address, “housing facility

Kerklaan" should be annotated as care institute, as we would replace this with the name of another care institute. The final annotation should look like this:

Example E.2. *In <DATE 2015> is hij met de andere cliënten verhuisd naar de <CARE-INSTITUTE woonvoorziening Kerklaan> in <ADDRESS Bennebroek>.*

[Translated] *In <DATE 2015> he moved with other clients to the <CARE-INSTITUTE housing facility Kerklaan> in <ADDRESS Bennebroek>.*

E.2 Example Annotations per Category

Table 12 provides example annotations for each of the PHI categories that were distributed to each annotator alongside with the instructions.

Category	Examples	Exclude from Annotation
Name	“Bart van der Boor”, “Boor, van der”, “B. Boor”, “Anne FP Jansen”	Titles (Dhr., Mw., Dr., etc.)
Initials	J.F., JF.	Titles (see above)
Profession	“stratenmaker”, “programmeur”, “militaire dienst”	Professions of medical staff
Hospital	“Universitair Medisch Centrum”, “UMC”	
Care Institute	“Cromhoff”, “Mgr. Bekkershuis”	
Organization	“Ikea”, “de Vink”, “de Efteling”	Generic names: “bouwmarkt”
Address	“Van Meeuwenstraat 2, 1234AB Town”, “Den Bosch”, “Nieuw-Zeeland”	
Internal Location	“8.1”	Generic locations: “surgery room”
Age	“2 jaar en 4 maanden”, “78”	
Date	“15-02-19”, “2019”, “vrijdag”, “Zomer ’02”, “herfstvakantie”, “koningsdag”,	Time of day (14:27)
ID	IBAN, license plate, employee number	
SSN/BSN	Burgerservicenummer (BSN)	

Table 12: Examples provided to the annotators alongside with the annotation instructions.