An Interactive De-Identification-System

Katrin Tomanek[1], **Philipp Daumke**[1], **Frank Enders**[1], **Jens Huber**[1], **Katharina Theres**[2] **and Marcel Müller**[2]
[1]Averbis GmbH, Freiburg/Germany
http://www.averbis.com
firstname.lastname@averbis.com
[2]Universitäts-Hautklinik, Freiburg/Germany
katharina.theres@gmx.de
marcel.mueller@uniklinik-freiburg.de

Abstract

We present a system for de-identification of unstructured clinical records. De-identification is performed semi-automatically in an interactive manner where the system suggests phrases of identifying information which need then be reviewed and verified by a human. The combination of automatic methods and manual approval ensures a high level of privacy and data security on the one hand and high throughput rates on the other hand.

1 Introduction

Clinical records contain plenty information highly valuable for clinical and operational research. Using text mining techniques, information hidden in *unstructured* records can be revealed and made accessible for data analysis. But since unstructured clinical records also contain plenty of private health information (PHI), such records are only accessible by a limited audience authorized to know the patient's identity. A prerequisite for access to clinical records outside of hospitals is their de-identification, i.e., the removal (or replacement) of all PHI phrases. De-identification is a labor-intensive tasks which constitutes a major bottleneck in the application of text mining techniques on clinical data.

We here present a system for *semi-automatic* de-identification of unstructured clinical records. The system is ready-to-use in a real-life clinical setting. It offers an intuitive graphical user interface, different de-identification projects can be managed and the system takes legal aspects when dealing with sensible clinical data in consideration by supporting different user roles and an explicit approval mechanism. Moreover, the complete de-identification workflow is supported, including import of different document types, annotation of PHI phrases, and export of de-identified texts to different formats. To the best of our knowledge, no comparable system is currently available. However, several algorithms to automatically find PHI phrases in text have been proposed (Meystre et al., 2010). Moreover, our system shared some features with other annotation tools (South et al., 2012) with the difference that it has been optimized for the task of de-identification in a clinical setting.

Automatic de-identification of structured records is rather straight-forward because critical data is only present in special fields. If unstructured data is in focus, de-identification is much harder and error-prone. State-of-the-art methods for automatic de-identification of unstructured data have detection rates of about 97% F1-score (Uzuner et al., 2007; Wellner et al., 2007). These results have been achieved on a relatively homogeneous set of records, i.e., medical discharge summaries from one single institution. However, considering that huge number of different document types and especially their variability across but also even within institutions, it is unclear which detection rates may be obtained in real-world applications.

De-identification of clinical records needs to be near-perfect when data should be accessible outside of hospitals. In consequence, we decided for an interactive de-identification approach to meet this high requirements: Our De-ID system automatically pre-annotates PHI phrases and then requires a human to

reviews these. The underlying intuition is that high throughput rates are achieved and de-identification can be performed much more efficient than in a completely manual scenario. The system can be applied to any type of clinical records. Preannotation is generic and record type-independent and a self-learning component quickly adapts to the document type and PHI elements in focus.

2 De-Identification Workflow

Our De-ID system supports the full workflow of de-identifying clinical records, starting with the import of "raw" clinical records as obtained from the hospital's IT system, annotation of PHI phrases, their replacement by placeholders or pseudonyms, and finally export of de-identified records in different formats for dissemination. Moreover, it offers tools for quality control and training of annotators.

2.1 Data Import

For application in day-by-day operations in medical institutions, the data formats as used in the local IT system should be supported to facilitate data exchange. Currently, our De-ID system supports different import formats including plain text, several types of the HL7 message standard, as well as some clinic-specific formats. The system can be easily extended to support other formats. If available, metadata is also imported and used for automatic preannotation of the documents (see Section 3).

2.2 Annotation of PHI Phrases

Figure 1 shows the PHI annotation workbench. Once a project has been selected, a list of all documents along with their processing status and the name of the last editor is available in the left panel (vertical tab *Document List*). Table 1 lists all processing status values along with a short description and the actions allowed for each status. Automatic preannotation, for instance, is only performed on *original* documents, editing the PHI phrase annotations is only possible for documents with status *original* or *in progress*, and only *approved* documents can be exported.

By double-click, a document is opened for annotation and its content is shown in the middle panel with annotated PHI phrases highlighted. Annotations can be added by marking the text passage and then selecting the respective PHI type (e.g., *location*) from a context menu or by using short-keys. The vertical tab *Document Information* on the left panel shows the document's status, a comment field and buttons to store the document and browse through the project's document collection. On the right panel, the document's metadata is shown. Also, a list of all annotated PHIs phrases can be shown (vertical tab *Protected Health Information*). For each PHI phrase, a confidence score and an annotation indicator, indicating how this phrase was annotated, is shown: PHI phrases can be either manually annotated (indicator *Manual*), or marked by one of the preannotation components (*Metadata*, *Regex*, and *Learn*; see Section 3). Annotations can be sorted by annotation indicator or confidence score to quickly spot problematic or insecure ones.

Along with the document's status, also the ID of the last editing user is stored. Users of role *annotator* usually perform PHI phrase annotation, while only users assigned the role *approver* can mark documents as *approved*. From a legal perspective, such an approval mechanism is crucial when documents shall be exported and used outside the hospital.

2.3 Export of De-Identified Data

After PHI phrases have been annotated, *approved* documents can be exported in a de-identified version where PHI phrases are replaced. At present, the system supports a simple replacement strategy, where all PHI phrases are replaced by a constant placeholder (e.g., "XXX"). Documents can be exported into different formats. When, e.g., the HL7 standard is used, the anonymized text can be written back to the original HL7 envelope.

2.4 Quality Control

The system offers tools for quality control in terms of inter-annotator agreement (IAA). Therefore, annotations of two projects can be compared. One of these project is defined as the gold standard, the other one as the project to be evaluated. IAA can be measured on token- or phrase-level in terms of recall, precision and F1-score. To further analyze disagreement, one can download a list of false negative and false positive annotations. Figure 2 shows results of an IAA test.

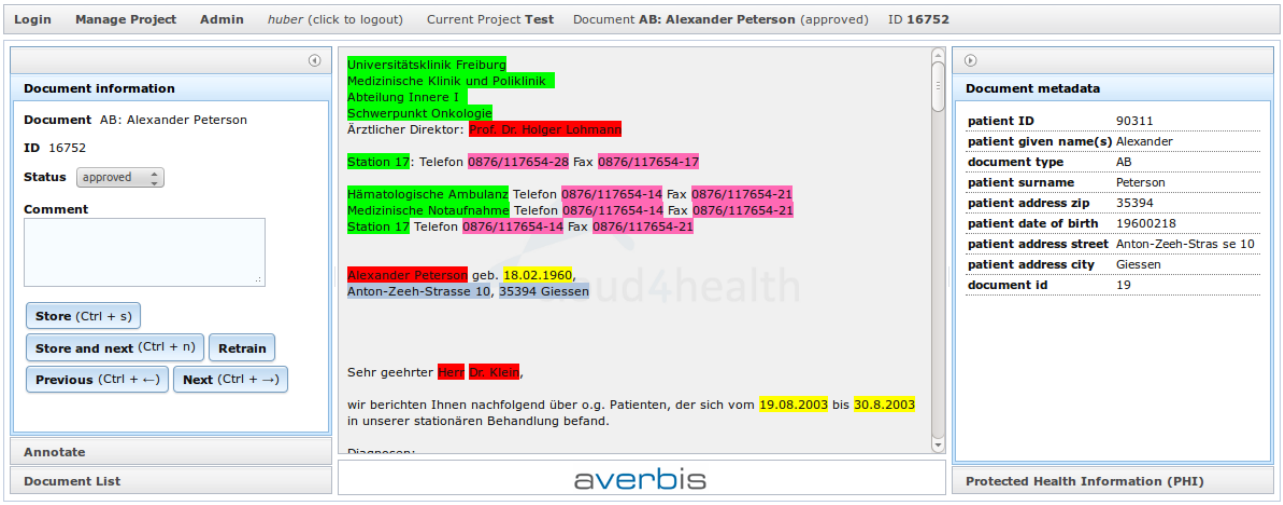

Figure 1: Annotation workbench

status	description	preannotate	annotate	export
original	no changes made / annotations contained	+	+	-
in progress	changes have been made to document	-	+	-
finalized	PHI phrase annotation finished; waiting for approval	-	-	-
approved	approver user has approved PHI annotations	-	-	+
rejected	rejected by approver, annotator must correct PHI annotation	-	+	-

Table 1: Document processing status values

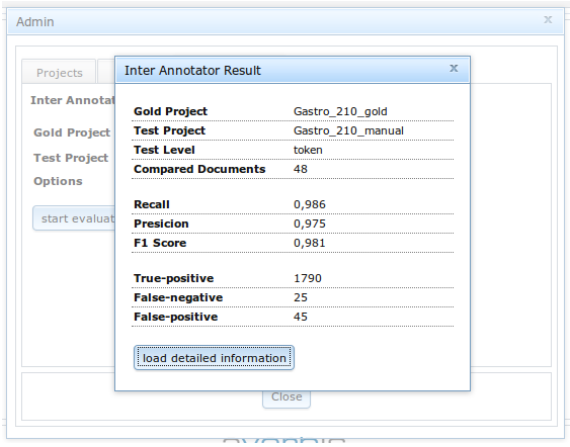

Figure 2: Calculation of Inter-Annotator-Agreement

3 Interactive De-Identification

Our system's automatic preannotation procedure consists of three levels, including metadata-matching, rule-based tagging as well as a component based on statistical machine learning (ML). While it has been shown that ML-based PHI recognition outperforms rule- or dictionary-based approaches (Uzuner et al., 2007), ML requires training material for the specific domain, genre and PHI types of interest. Since such data is not readily available in most practical settings, our approach to preannotation is *hybrid* and *self-learning*.[1] Its hybridity allows for the system to be able to detect PHI phrases based on metadata-matching and rules when there is no or insufficient training data available (usually when a de-identification project has just started).

Whenever PHI phrase annotation of a document is finished, the document is automatically added to the training set from which ML-based system is trained in a background process. The ML-based preannotation improves as more documents are annotated and provides increasingly better support to the human annotator. From our experience, even a few annotated documents often suffice to learn a model which can fill many of the recall-gaps of the metadata-matcher and the rule-based tagger.

[1]There is high variability of document type, formats and writing style across and even within medical institutions. Thus a PHI model learned from one set of medical records will most likely be sub-optimal when re-used in a different context.

Metadata-Matching When metadata such as patient name, contact information, date of birth etc are available for a document (this is, e.g., the case when HL7 messages are imported), the metadata-matching component performs exact string matching to identify PHI phrases referring to these metadata elements. We refrain from fuzzy matching for the sake of precision. To increase recall, we added a set of variations and combinations of metadata elements (e.g. for street names, date formats or combinations of given and surname).

Rule-based Tagging Many PHIs can also be detected with high precision based on predefined rules. This includes for example mentions of dates, person names when combined with certain titles ("Herr Maier" or "Prof. Smith"), dates, telephone numbers, zip codes together with place names, and names of hospitals and divisions ("Klinik für Innere Medizin") when mentioned together with indicator words. A goal in developing these rules was them to be very precise (potentially at the cost of recall) and general so they would be applicable to different record types from different medical domains.[2]

ML-based Tagging Our system applies Conditional Random Fields (Lafferty et al., 2001), a framework for sequence tagging which has been successfully applied before fore Named Entity Recognition (NER) and de-identification (Wellner et al., 2007). The document os modeled as a sequence of words where each word is assigned one PHI type. We rely on standard NER features extended slightly to represent characteristics of clinical records (e.g., length of a sentence, position of a word within the record to reflect header/footer properties). We aimed at building a component which is general enough to work well on different types of records instead of being optimized for a single type. For example, evaluated on the data of the I2B2 de-identification challenge (Uzuner et al., 2007), our ML system achieved an F1-score of about 94% out of the box without any domain- or language-specific adaptions.

[2]We are aware that in the de-identification task, recall takes precedence over precision. However, rules which produce many false positives will increase the annotation effort in an interactive system (manual removal of incorrect PHI phrases). To increase recall in specific scenarios, our system comes with a self-learning component.

4 System Architecture

The De-ID system consists of three main architectural components: *a)* the GUI for annotation and management or projects, *b)* a data store, and *(c)* the NLP framework for document preprocessing and preannotation. The GUI is a web-application allowing remote access to the system through a web-browser. This avoids the need to transfer sensitive data in its entirety to the annotators. User and project information as well as imported documents and their annotations are stored on the server in a database. The UIMA framework[3] is used for document processing. When imported, documents are sent through a UIMA pre-processing pipeline consisting of components for sentence splitting, tokenization and shallow syntactic processing. For pre-annotation, documents are sent through a pipeline consisting of the preannotation components. The De-ID system comes with default components for German documents. Single components can be easily reconfigured or exchanged to meet language- or application-specific requirements.

5 Summary

We have presented an interactive, self-learning system for de-identification. The system is designed to be as generic as possible allowing it to work well on different types of clinical records.

We are currently running studies on different types of clinical records in collaboration with German hospitals to test *a)* the usability of the system in a clinical setting, *b)* our hypothesis that interactive de-identification is more efficient than annotation without preannotation, *c)* how much data the ML-based component needs for reliable predictions, and *d)* how well preannotation works in different settings which different types of records.

In the second version of the system, we will improve the replacement mechanism. In a first step, PHI phrases will be semantically interpreted and normalized so that in a second step type-dependent replacements can be made.[4]

[3]`http://uima.apache.org/`

[4]A PHI phrase of type date may be subdivided into day, month and year and normalized to numbers (example: *1. Mai 2012* will be *d=1, m=5, y=2012*). The normalized date could then be replaced by a coarser date, e.g. the quarter (2/2012).

References

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc.

S. Meystre, F. Friedlin, B. South, S. Shen, and M. Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):70.

Brett South, Shuying Shen, Jianwei Leng, Tyler Forbush, Scott DuVall, and Wendy Chapman. 2012. A prototype tool set to support machine-assisted annotation. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139. Association for Computational Linguistics.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Viewpoint paper: Evaluating the state-of-the-art in automatic de-identification. *JAMIA*, 14(5):550–563.

Ben Wellner, Matt Huyck, Scott A. Mardis, John S. Aberdeen, Alexander A. Morgan, Leonid Peshkin, Alexander S. Yeh, Janet Hitzeman, and Lynette Hirschman. 2007. Research paper: Rapidly retargetable approaches to de-identification in medical records. *JAMIA*, 14(5):564–573.