A Hybrid Stepwise Approach for De-identifying Person Names in Clinical Documents

Oscar Ferrández[1,2], Brett R. South[1,2], Shuying Shen[1,2], Stéphane M. Meystre[1,2]
[1] Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[2] IDEAS Center SLCVA Healthcare System, Salt Lake City, Utah, USA
oscar.ferrandez@utah.edu,
{brett.south,shuying.shen,stephane.meystre}@hsc.utah.edu

Abstract

As Electronic Health Records are growing exponentially along with large quantities of unstructured clinical information that could be used for research purposes, protecting patient privacy becomes a challenge that needs to be met. In this paper, we present a novel hybrid system designed to improve the current strategies used for person names de-identification. To overcome this task, our system comprises several components designed to accomplish two separate goals: 1) achieve the highest recall (no patient data can be exposed); and 2) create methods to filter out false positives. As a result, our system reached 92.6% F_2-measure when de-identifying person names in Veteran's Health Administration clinical notes, and considerably outperformed other existing "out-of-the-box" de-identification or named entity recognition systems.

1 Introduction

Electronic Healthcare Records are invaluable resources for clinical research, however they contain highly sensitive Protected Health Information (PHI) that must remain confidential. In the United States, patient confidentiality is regulated by the Health Insurance Portability and Accountability Act (HIPAA). To share and use clinical documents for research purposes without patient consent, HIPAA requires prior removal of PHI. More specifically, the HIPAA "Safe Harbor"[1] determines 18 PHI categories that have to be obscured in order to consider clinical data de-identified.

An ideal de-identification system should recognize PHI accurately, but also preserve relevant non-PHI clinical data, so that clinical records can later be used for various clinical research tasks.

Of the 18 categories of PHI listed by HIPAA, one of the most sensitive is patient names, and all person names in general. Failure to de-identify such PHI involves a high risk of re-identification, and jeopardizes patient privacy.

In this paper, we describe our effort to satisfactorily de-identify person names in Veteran's Health Administration (VHA) clinical documents. We propose improvements in person names de-identification with a pipeline of processes tailored to the idiosyncrasies of clinical documents. This effort was realized in the context of the development of a best-of-breed clinical text de-identification system (nicknamed "BoB"), which will be released as an open source software package, and it started with the implementation and evaluation of several existing de-identification and Named Entity Recognition (NER) systems recognizing person names. We then devised a novel methodology to better tackle this task and improve performance.

2 Background and related work

In many aspects de-identification resembles traditional NER tasks (Grishman and Sundheim, 1996). NER involves detecting entities such as person names, locations, and organizations. Consequently, given the similar entities targeted by both tasks, NER systems can be relevant to de-identify documents. However, most named entity recognizers were developed for newswire articles, and not for clinical narratives. Clinical records are character-

[1] GPO US: 45 C.F.R. § 164 Security and Privacy. http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html Further details about the 18 HIPAA Safe Harbor PHI identifiers can be also found in (Meystre et al., 2010).

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 65–72,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

ized by fragmented and incomplete utterances, lack of punctuation marks and formatting, as well as domain specific language. These complications, in addition to the fact that some entities can appear both as PHI and non-PHI in the same document (e.g., "Mr. Epley" vs. "the Epley maneuver"), make clinical text de-identification a challenging task. Therefore, although person names de-identification is essentially NER, the unique characteristics of clinical texts make it more interesting and challenging than recognizing names in news articles, which also enhance the motivation for this study.

Several different approaches were proposed to deal with de-identification of clinical documents, and for named entity recognition of person names. These approaches are mainly focused on either pattern matching techniques, or statistical methods (Meystre et al., 2010), as exemplified below.

Beckwith et al. (2006) developed a de-identification system for pathology reports. This system implemented some patterns to detect dates, locations, and ID numbers, as well as a database of proper names and well-known markers such as 'Mr.' and 'PhD' to find person names.

Friedlin and McDonald (2008) described the Medical De-identification System (MeDS). It used a combination of methods including heuristics, pattern matching, and dictionary lookups to identify PHI. Pattern matching through regular expressions was used to detect numerical identifiers, dates, addresses, ages, etc.; while for names, MeDS used lists of proper names, common usage words and predictive markers, as well as a text string nearness algorithm to deal with typographical errors.

Neamatullah et al. (2008) proposed another rule-based de-identification approach focused on pattern matching via dictionary lookups, regular expressions and context checks heuristics denoting PHI. Dictionaries made up of ambiguous names and locations that could also be non-PHI, as well as dictionaries of common words were used by this system to disambiguate PHI terms.

Other de-identification systems such as (Aberdeen et al., 2010; Gardner and Xiong, 2009) use machine learning algorithms to train models and predict new annotations. The key aspect of these systems is the selection of the learning algorithm and features. Both (Aberdeen et al., 2010) and (Gardner and Xiong, 2009) use an implementation of Conditional Random Fields (CRF) and a set of learning features based on the morphology of the terms and their context. One disadvantage of these systems is the need for large amounts of annotated training examples.

As mentioned previously, for detecting person names, we could also use traditional newswire-trained NER systems. NER has long been studied by the research community and many different approaches have been developed (Tjong Kim Sang and De Meulder, 2003; Doddington et al., 2004). One successful and freely available named entity recognizer is the Stanford NER system (Finkel et al., 2005), which provides an implementation of linear chain CRF sequence models, coupled with well-engineered feature extractors for NER, and trained with newswire documents.

3 Methods

As already mentioned, we first selected and ran several existing de-identification and NER systems detecting person names in our clinical documents. Afterwards, we devised and present here a novel pipeline of processes designed to improve the PHI recognition task.

3.1 Existing de-identification and NER systems

Five available de-identification systems, as well as one newswire-trained named entity recognizer, were selected for an "out-of-the-box" evaluation. The aim of this evaluation was to compare the performance of the various methods and resources when de-identifying person names in our clinical documents.

We included three rule-based de-identification approaches:

- HMS Scrubber (Beckwith et al., 2006);
- MeDS (Friedlin and McDonald, 2008); and
- MIT deid system (Neamatullah et al., 2008).

And two systems based on machine learning classifiers:

- The MITRE Identification Scrubber Toolkit (MIST) (Aberdeen et al., 2010); and
- The Health Information DE-identification (HIDE) system (Gardner and Xiong, 2009).

Regarding NER systems, we chose the Stanford NER system (Finkel et al., 2005), which has reported successful results when detecting person names. These systems were described in Section 2, when we presented related work.

3.2 Our best-of-breed approach

Our names de-identification approach consists of a novel pipeline of processes designed to improve the current strategies for person names de-identification. This system is being developed as an Apache UIMA[2] pipeline, with two main goals:

1) Obtain the highest recall (i.e., sensitivity), regardless of the impact on precision; and
2) Improve overall precision by filtering out the false positives produced previously.

These goals correspond to the implementation of the main components of our system. When we tested existing systems (we will present results for these systems in Table 1), we observed that recall was better addressed by rule-based approaches, while precision was higher applying machine learning-based algorithms. We therefore used this knowledge for the design of our system: goal#1 is then accomplished mainly using rule-based techniques, and goal#2 implementing machine learning-based approaches.

Moreover, recall is of paramount importance in de-identification (patient PHI cannot be disclosed). And this was also a reason that motivated us to first focus on achieving high recall, and filtering out false positives afterwards as a separate procedure.

Unlike other de-identification and NER systems that tackle the classification problem from one perspective (i.e., rule-based or machine learning-based) or from a limited combined approach (e.g., learning features extracted using regular expressions), the design of our system allows us to take advantage of the strong points of both techniques separately. And more importantly, our classifiers for filtering out false-positives (goal#2) are trained using correct and incorrect annotations derived from previous modules implemented in goal#1. Thus, they do not predict if every token in the document is or belongs to a PHI identifier, they instead decide if an actual annotation is a false or true positive. This design makes our classifiers better with less learning examples, which is a restriction we have to deal with, and it also allows us to create methods that can be only focused on maximizing recall regardless of the amount of false-positives introduced (goal#1). To the best of our knowledge, this perspective has not been exploited before, and as we will show in the evaluation section, it empirically demonstrates more robustness than previous approaches.

The design of our system integrates different components described below. Figure 1 depicts an overview of our system's architecture and workflow.

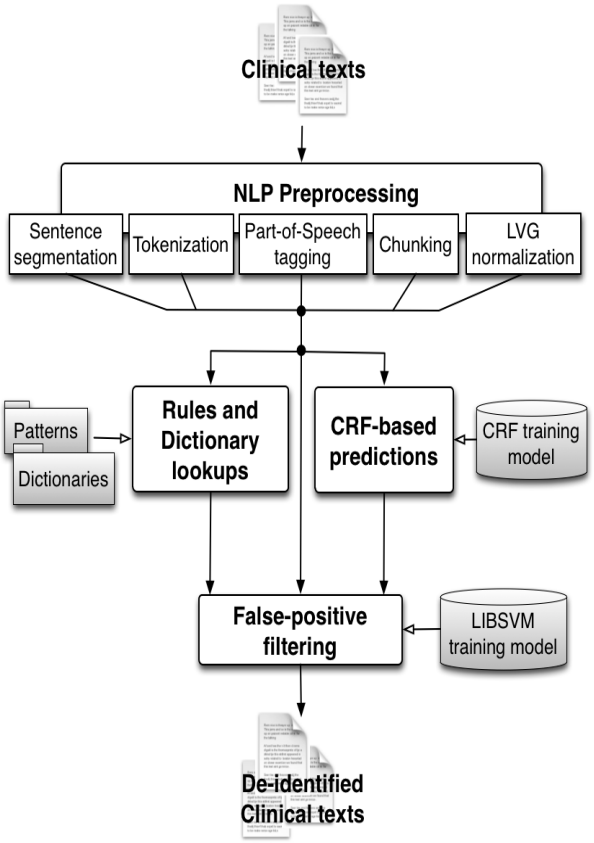

Figure 1. System's architecture.

3.2.1 NLP preprocessing steps

This NLP preprocessing prepares the input for the main components of our system. It includes sentence segmentation, tokenization, part-of-speech tagging, chunking, and word normalization based

[2] http://uima.apache.org/

on Lexical Variant Generation (LVG)[3]. The output of this preprocessing will be used by subsequent pattern matching techniques and features for machine learning algorithms. For these processes, we adapted several cTAKES (Savova et al., 2010) components.

3.2.2 Rules and dictionary lookups

We created a pattern matching component supported by contextual keyword searches (e.g., "Dr.", "Mr.", "M.D.", "R.N.", "L.C.S.W."), dictionaries of person names[4], and a simple disambiguation procedure based on a list of common words and the capitalization of the entity. We adapted some of the techniques implemented in (Beckwith et al., 2006; Friedlin and McDonald, 2008; Neamatullah et al., 2008) to our documents, and developed new patterns. For dictionary lookups, we used Lucene[5] indexing, experimenting with keyword and fuzzy dictionary searches. Each word token is compared with our indexed dictionary of names (last and first names from the 1990 US Census[4]), considering all matches as candidate name annotations. However, candidates that also match with an entry in our dictionary of common words[6] and do not contain an initial capital letter are discarded from this set of candidate name annotations.

With this component, we attempt to maximize recall, even if precision is altered.

3.2.3 CRF-based predictions

To further enhance recall, we created another component based on CRF models. We incorporated this component in our system considering that machine learning classifiers are more generalizable and can detect instances of names that are not supported by our rules or dictionaries. Therefore, although we knew the individual results of a CRF classifier at this level were not enough for de-identification, at this point our main concern is to obtain the highest recall. Thus, adding a machine learning classifier into this level we could help the system predicting the PHI formats and instances that could not be covered by our patterns and dictionaries.

To develop this component, we used the CRF classifier implementation provided by the Stanford NLP group[7]. We carried out a feature selection procedure using greedy forward selection. It provided us with the best learning feature set, which consisted of: the target word, 2-grams of letters, position in the document, part-of-speech tag, lemma, widely-used word-shape features (e.g., initial capitals, all capitals, digits inside, etc.), features from dictionaries of names and common words, a 2-word context window, and combinations of words, word-shapes and part-of-speech tags of the word and its local context.

The learning features considered before and after the selection procedure are shown in Table 1.

3.2.4 False-positive filtering

The two previous components' objective is maximal recall, producing numerous false positives. The last component of our pipeline was therefore designed to filter out these false positives and consequently increase overall precision. We built a machine learning classifier for this task, based on LIBSVM (Chang and Lin, 2001), a library for Support Vector Machines (SVM), with the RBF (Radial Basis Function) kernel. We then trained this classifier with reference standard text annotations, as well as the correct and incorrect annotations made by the previous components. We used our training document set (section 4.1) for this purpose.

Features for the LIBSVM machine learning model were: the LVG normalized form of the target annotation, three words before and after, part-of-speech tags of the words within the annotation and the local context, number of tokens within the annotation, position in the document, 40 orthographic features (denoting capitals, digits, special characters, etc.), features from dictionaries of names and common words, and the previous strategy used to make the annotation (i.e., rules, dictionary lookups or CRF-based predictions).

[3] http://lexsrv2.nlm.nih.gov/LexSysGroup/Projects/lvg/current/web/index.html

[4] Frequently Occurring Names from the 1990 Census. http://www.census.gov/genealogy/names.

[5] http://lucene.apache.org/java/docs/index.html

[6] We used the dictionary of common words from Neamatullah et al. (2008).

[7] http://nlp.stanford.edu/software/corenlp.shtml

Feature	Description	Selected*
target word	The word to classify as person name	Yes
2-grams of letters	Features from the 2-grams of letters from the word	Yes
3-grams of letters	Features from the 3-grams of letters from the word	No
4-grams of letters	Features from the 4-grams of letters from the word	No
lowercase n-grams	Features from the n-grams of letters from the word in lowercase (considering 2-, 3-, and 4-grams separately)	No
position	Position of the word within a sentence	Yes
PoS	Part-of-speech tag of the word	Yes
lemma	Lemma of the word	Yes
word shape	Initial capital	Yes
	All capitals	
	Mix of uppercase and lowercase letters	
	Digits inside	
	All digits	
	Has dash	
	End dash	
	Alpha-numeric	
	Numeric-alpha (starts with a number)	
	Contains punctuation mark	
dictionaries	Does the word match with an entry of the dictionary of names?	Yes
	Does the word match with an entry of the dictionary of common words?	Yes
2-word window	The two preceding and following words in the context	Yes
3-,4-,5-word window	The three, four and five preceding and following words in the context	No
word-pairs	Combinations of the word and the next and previous words in the context window, preserving direction but not position (considering separate features for the different combinations of the context and the target word)	No
titles	Match the word against a list of name titles (Mr, Mrs, etc.)	No
lemma_context	Lemma of the words inside the contextual window	No
PoS_context	Individual features from the part-of-speech tags of the contextual window	Yes
PoS_sequence	Sequence of the part-of-speech tags of the 2-word contextual window and the target word	Yes
word_shape_context	Word shape features of the contextual window	Yes
word-tag	Combination of the word and part-of-speech	No

Table 1. Set of learning features for the CRF-based prediction module. (* = selected in the best learning features set)

4 Evaluation and discussion

Our evaluation consists of: 1) "out-of-the-box" evaluation of the systems presented in Section 3.1; and 2) evaluation of the performance of our person names de-identification pipeline.

4.1 Data

We manually annotated all person names (including patients, relatives, health care providers, and other persons) in a corpus of various types of Veteran's Health Administration (VHA) clinical notes. These notes were selected using a stratified random sampling approach with documents longer than 500 words. Then, the 100 most frequent VHA note types were used as strata for sampling, and the

same number of notes was randomly selected in each stratum. Two reviewers independently annotated each document, a third reviewer adjudicated their disagreements, and a fourth reviewer eventually examined ambiguous and difficult adjudicated cases.

The evaluation corpus presented here comprises a subset of 275 VHA clinical notes from the aforementioned corpus. For training, 225 notes were randomly selected (contained 748 person name annotations), and the remaining 50 notes (with 422 name annotations) were used for testing the systems.

4.2 Experiments and results

We present results in terms of precision, recall and F-measure (harmonic mean of recall and precision). We used a weight of 2 when calculating the F_2-measure giving recall more (twice) importance than precision (Jurafsky and Martin, 2009). This reflects our emphasis on recall for de-identification. To our understanding, due to legal and privacy issues, a good de-identification system should be tailored to prioritize recall, and consequently patient confidentiality. It is not the scope of this paper to judge or modify the development design adopted by other de-identification systems.

Moreover, we considered correct predictions at least overlapping with the entire PHI annotation in the reference standard (i.e., exact match with the reference annotation, or more than the exact match). We can therefore assure complete redaction of PHI.

Table 2 illustrates "out-of-the-box" evaluation results of the systems described in Section 3.1. For this evaluation, we trained MIST and HIDE with our 225 notes training corpus, while the Stanford NER was run using the trained models available with its distribution[8]. Testing was realized using our 50 notes testing corpus.

Table 3 shows the performance of our names de-identification approach. We provide results for different configurations of our pipeline:

- **Rules & Dictionaries**. Results of the rules and dictionary lookups component described in Section 3.2.2, in this case using a keyword-search strategy for dictionary lookups.
- **R&D with fuzzy searches**. Results from the rules and dictionary lookups component using Lucene's Fuzzy Query engine for dictionary searches. It implements a fuzzy search based on the Levenshtein (edit distance) algorithm[9] (Levenshtein, 1966), which has to surpass a similarity threshold in order to produce a match. We carried out a greedy search on the training corpus for the best similarity threshold. We found 0.74 to be the best threshold.
- **CRF-based w/FS**. The CRF-based predictions component results after selecting the best set of features (see Section 3.2.3). The CRF classifier was trained using our 225-document training corpus.
- **R&D + CRF w/FS**. The cumulative results from the rules and dictionary lookups (not implementing fuzzy dictionary searches) and the CRF-based predictions components.
- **R&D + CRF w/FS + FP-filtering**. Includes all components together, adding the false-positive filtering component (Section 3.2.4) at the end of the pipeline. The SVM model for this last component was created using our training corpus.

System	Prec.	Rec.	F_2
HMS Scrubber	0.150	0.675	0.397
MeDS	0.149	0.768	0.419
MIT deid	0.636	0.893	0.826
MIST	0.865	0.319	0.356
HIDE	0.975	0.376	0.429
Stanford NER	0.692	0.723	0.716

Table 2. "Out-of-the-box" evaluation of existing de-identification and NER systems (Prec.=precision; Rec.=recall; F_2= F_2-measure).

System	Prec.	Rec.	F_2
Rules & Dictionaries	0.360	0.962	0.721
R&D + fuzzy	0.171	0.969	0.502
CRF-based w/FS	0.979	0.874	0.893
R&D + CRF w/FS	0.360	0.988	0.732
R&D + CRF w/FS + FP-filtering	0.774	0.974	**0.926**

Table 3. Cumulative results of our pipeline of processes.

[8] Further details about these models can be found at http://nlp.stanford.edu/software/CRF-NER.shtml

[9] http://www.merriampark.com/ld.htm

4.3 Analysis

Our novel names de-identification pipeline significantly outperforms all other systems we evaluated "out-of-the-box" or trained with our VHA notes corpus.

Among the five existing systems we evaluated (Table 1), only one achieved noteworthy recall around 89%. However, none of them obtained any remarkable F_2-measure. Most errors produced by the pattern matching systems (i.e., HMS Scrubber, MeDS, and MIT deid system) were due to false positive annotations of medical eponyms (e.g., "Achilles", "Guyon", etc.), as well as acronyms denoting medical facilities (e.g., "ER" and "HCS"). The false negatives consisted of ambiguous person names (e.g., "Bill" and "Chase"), some formats not covered by the patterns (e.g., "[LastName], [FirstName] [Initial]"), and a few names not found in the dictionaries.

Among machine learning-based systems, the two de-identification applications (i.e., MIST and HIDE) obtained good precision, but quite low recall. The size of our training corpus was somewhat limited, and these results probably indicate a need for more sophisticated learning features, as well as feature selection procedures (rather than using the "out-of-the-box" feature specification that comes with these systems) for better performance. With improved learning features, we could mitigate the relative lack of training examples. Interestingly, the NER system, which was trained on newswire documents, performed even better than some de-identification systems, although a need for improvement is still present.

We acknowledge that the comparison with Stanford NER is not completely fair due to the different source of documents used for training. However, we considered it interesting information, and although clinical notes contain characteristics not present in newswire corpora, they also have similarities regarding person names (e.g., titles "Mr.", "Dr.", "PhD", part-of-speech, verb tenses). Therefore, we think that only for names recognition, a newswire trained NER can provide interesting results, and this was actually what we observed.

Table 2 points out that the combination of our components produces successful cumulative results. Using the training corpus to create a simple component made up of rules, dictionary lookups, and few heuristics for disambiguation allowed for recall values of 0.96. This demonstrates the need to adapt these techniques to the target documents, instead of employing systems "out-of-the-box".

Our experiments with fuzzy dictionary lookups did not allow for a significant increase in recall, but caused a decrease in precision (-19%). It suggests that there was no need for considering person name misspellings.

The component based on CRF predictions alone achieved good performance, especially in precision. It obtained the best F_2-measure (0.89), clearly higher than the other "out-of-the-box" systems based on CRF models. It proves that selecting suitable learning features mitigates to some extent the scarcity of training examples.

Our next experiment combined the rules and dictionaries and CRF components. It improved the overall recall to about 0.99, which means that CRF-based predictions recognized some person names that were missed by our pattern matching components, but didn't increase the precision. We reached here our first goal of high recall or sensitivity.

Finally, we added the false-positive filtering component to our system. This component was able to filter out 622 (84%) false positives from a total of 742, improving the precision to 0.77 (+41%); but also causing a slight decrease in recall (-1.4%). This application of our pipeline was successful, reaching an F_2-mesure of 0.93, and was an effective way of training the SVM model for false-positives filtering.

5 Conclusions

We designed and evaluated a novel person names de-identification system with VHA clinical documents. We also presented an "out-of-the-box" evaluation of several available de-identification and NER systems; all of them were surpassed by our approach.

With our proposal, we showed that it is possible to improve the recognition of person names in clinical records, even when the corpus for training machine learning classifiers is limited. Furthermore, the workflow of our pipeline allowed us to tackle the de-identification task from an intuitive but powerful perspective, i.e. facing the achievement of high recall and precision as two separate goals implementing specific techniques and components.

Packaging this two-step procedure as a bootstrapping learning or adding the rules to define learning features would not allow us to use the qualities of the R&D and CRF components (i.e., obtain the highest recall by any means). Moreover, considering the small size of our manually annotated examples, these approaches would not work much better than existing systems.

As future efforts, we plan to improve the precision of the rules and dictionary lookups component by adding more sophisticated person names disambiguation procedures. Such procedures should deal with the peculiar formatting of clinical records as well as integrate enriched knowledge from biomedical resources. We also plan to evaluate the portability of our approach by using other sets of clinical documents, such as the 2006 i2b2 de-identification challenge corpus (Uzuner et al., 2007).

Acknowledgments

Funding provided by the Department of Veterans Affairs Health Services Research & Development Services Consortium for Healthcare Informatics Research grant (HIR 08-374).

References

John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. *The MITRE Identification Scrubber Toolkit: design, training, and assessment*. International journal of medical informatics, 79 (12) (December): 849-59.

Bruce A. Beckwith, Rajeshwarri Mahaadevan, Ulysses J. Balis, and Frank Kuo. 2006. *Development and evaluation of an open source software tool for deidentification of pathology reports*. BMC medical informatics and decision making, 6 (1) (January): 12.

Chih-Chung Chang and Chih-Jen Lin. (2001). *LIBSVM: a library for support vector machines*. Computer, 1-30.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. *The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation*. Proceedings of LREC 2004: 837-840.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. *Incorporating non-local information into information extraction systems by Gibbs sampling*. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics: 363-370.

F. Jeff Friedlin and Clement J. McDonald. 2008. *A software tool for removing patient identifying information from clinical documents*. Journal of the American Medical Informatics Association : JAMIA 15 (5) (January 1): 601-10.

James Gardner and Li Xiong. 2009. *An integrated framework for de-identifying unstructured medical data*. Data & Knowledge Engineering 68 (12) (December): 1441-1451.

Ralph Grishman and Beth Sundheim. 1996. *Message understanding conference-6: A brief history*. Proceedings of the 16th conference on Computational linguistics - Volume 1: 466-471. Association for Computational Linguistics, Copenhagen, Denmark.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall. Upper Saddle River, NJ, USA.

V.I. Levenshtein. 1966. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*. Soviet Physics - Doklady 10: 707–710.

Stephane M. Meystre, F. Jeffrey Friedlin, Brett R. South, Shuying Shen, and Matthew H. Samore. 2010. *Automatic de-identification of textual documents in the electronic health record: a review of recent research*. BMC medical research methodology 10 (1) (January): 70.

Ishna Neamatullah, Margaret M. Douglass, Li-wei H. Lehman, Andrew 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 medical informatics and decision making 8 (1) (January): 32.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. Journal of the American Medical Informatics Association : JAMIA 17 (5): 507-13.

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: 142-147.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. *Evaluating the State-of-the-Art in Automatic De-identification*. Journal of the American Medical Informatics Association : JAMIA 14(5):550-563.