De-Identification of PHI from Clincial Text Data

Christopher Meaney

Biostatistician
Department of Family and Community Medicine
University of Toronto

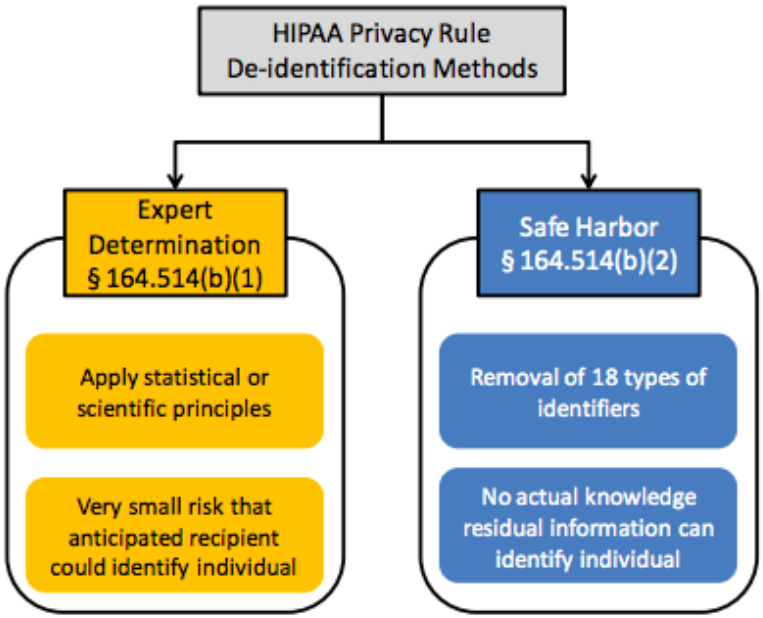
September 28, 2020

Rationale for De-identification

> The increasing adoption of health information technologies in the United States accelerates their potential to facilitate beneficial studies that combine large, complex data sets from multiple sources. The process of de-identification, by which identifiers are removed from the health information, mitigates privacy risks to individuals and thereby supports the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors.
>
> — Health Insurance and Portability and Protection Act (1996)

Goals for De-identification Systems

- Reduce re-identification risk, replacing PHI with realistic surrogates.
- Maintain medically relevant information and biomedical semantic structure from text data.
- Use privacy/security as a tool to enhance public trust in the secondary use of medical data.



Who Defines PHI and What to De-Identify?

- HIPAA (USA) specified two approaches to DEID: (1) expert determination, (2) safe-harbor.
- Other jurisdictions lack explicit PHI definitions (PHIPAA-Canada; GDPR-EU).
- Do we need more risk averse interpretations of HIPAA? Is HIPAA enough?

No.	PHI Type
1	Names
2	All geographic subdivisions smaller than a state
3	Dates
4	Telephone Numbers
5	Vehicle Identifiers
6	Fax Numbers
7	Device Identifiers and Serial Numbers
8	Emails
9	URLs
10	Social Security Numbers
11	Medical Record Numbers
12	IP Addresses
13	Biometric Identifiers
14	Health Plan Beneficiary Numbers
15	Full-face photographic images and any comparable images
16	Account Numbers
17	Certificate/license numbers
18	Any other unique identifying number, characteristic, or code.

Approaches to De-Identification of Medical Notes

- Complete redaction of PHI.
- Substitution with granular identification tokens.
- Substitution with realistic surrogate information.

Replacements for Example Note			
Original	Substituted	Identifier	Redacted
Dear Janine Keane, as we have discussed, I hereby send you the requested information about my patient, Julie Andrews. You can reach her via email (her address is jandrews@gmail.com) or via phone: 998 785 6756. Sincerely, Elijah Hunt, MD	Dear Rosie Copeland, as we have discussed, I hereby send you the requested information about my patient, Beatrice Burton. You can reach her via email (her address is lebewukak@halava.fi) or via phone: (836) 230-3149. Sincerely, Jayden Bush, MD	Dear <name>, as we have discussed, I hereby send you the requested information about my patient, <name>. You can reach her via email (her address is <email>) or via phone: <phone>. Sincerely, <name>, MD	Dear **********, as we have discussed, I hereby send you the requested information about my patient, **********. You can reach her via email (her address is **********) or via phone: **********. Sincerely, **********, MD

Methods for de-identification of PHI

- Rule based systems.
  - REGEX, pattern matching, dictionaries.
- Machine learning systems.
  - Conditional random fields popular.
  - Require heavy feature engineering.
- Hybrid ML/rule-based systems.
- Deep learning systems.
  - Little/no feature engineering required.
  - RNNs, bi-LSTMs, Transformers.

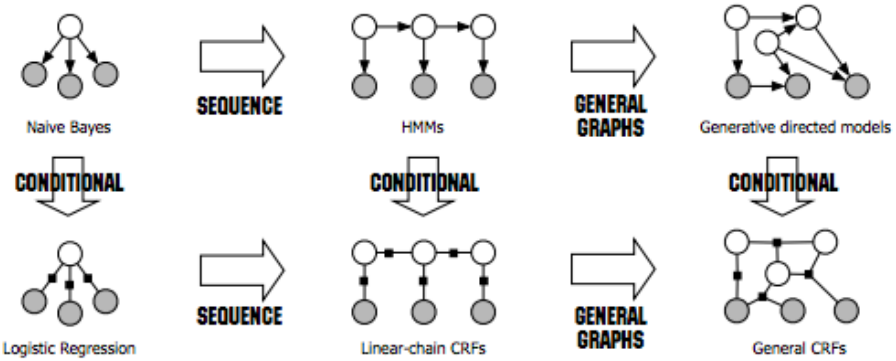


Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

Examples of Engineered Features

Bag-of-words: unigrams, bigrams and trigrams of words within a window of [-2, 2].
Part-of-speech (POS) tags: unigrams, bigrams and trigrams of POS tags within a window of $[-2, 2]$. The Stanford POS Tagger [42] was used for POS tagging.
Combinations of words and POS tags: combining current word with the unigrams, bigrams and trigrams of POS tags within a window of $[-1, 1]$, i.e. w_0p_{-1}, w_0p_0, w_0p_1, $w_0p_{-1}p_0$, $w_0p_0p_1$, $w_0p_{-1}p_1$, $w_0p_{-1}p_0p_1$, where w_0, p_{-1}, p_0 and p_1 denote current word, last, current and next POS tags respectively.
Sentence information: number of words in current sentence, whether there is an end mark at the end of current sentence such as '.', '?' and '!', whether there is any bracket unmatched in current sentence.
Affixes: prefixes and suffixes of length from 1 to 5.
Orthographical features: whether the word is upper case, contains uppercase characters, contains punctuation marks, contains digits, etc.
Word shapes: mapping any or consecutive uppercase character (s), lowercase character(s), digit(s) and other character(s) in current word to ' A', ' a', ' #' and ' -' respectively. For instance, the word shapes of "Hospital" are "Aaaaaaaa" and "Aa".
Section information: twenty-nine section headers (see the supplementary file) were collected manually such as "History of Present Illness"; we check which section current word belongs to.
General NER information: the Stanford Named Entity Recognizer [43] was used to generate the NER tags of current word, include: person, date, organization, location, and number tags, etc.
Word representation features: two types of word representation features generated by Brown clustering [44] and word2vec [45] on training sets and a large unlabeled MEDLINE corpus.
Dictionary features: four categories of localization dictionaries: COUNTRY, STATE, CITY and ZIP were collected from Internet, and each word was labeled with ' 0' or ' 1' by dictionary lookup.

Benchmark Data for DEID Research

- Uzuner (2007). Eval SOTA Auto-DEID.
  - Results i2b2 2006 DEID challenge.
  - Discharge summaries, Parters HC.
  - 889pt, 500k tokens (20k PHI).
- Neamatullah (2008). Auto-DEID EMRs.
  - Sample nursing notes MIMIC-II.
  - 169pt, 2k note. 330k tok (2k PHI).
- Stubbs (2014). Long Clin Narr DEID.
  - i2b2 2016 DEID challenge (PHC).
  - Longitudinal notes diab/cvd pats.
  - 300pt, 1k nt. 800k tok (30k PHI).
- Stubbs (2016). DEID Psych Adm Notes.
  - CEGS-NGRID DEID challenge.
  - Psychiatric intake notes Partners.
  - 1k records, 1.8M tok (35k PHI).
- Uzuner (2017). RNNs for DEID.
  - MIMIC-III ICU discharge note.
  - 1600pt, 3M tokens (61k PHI).

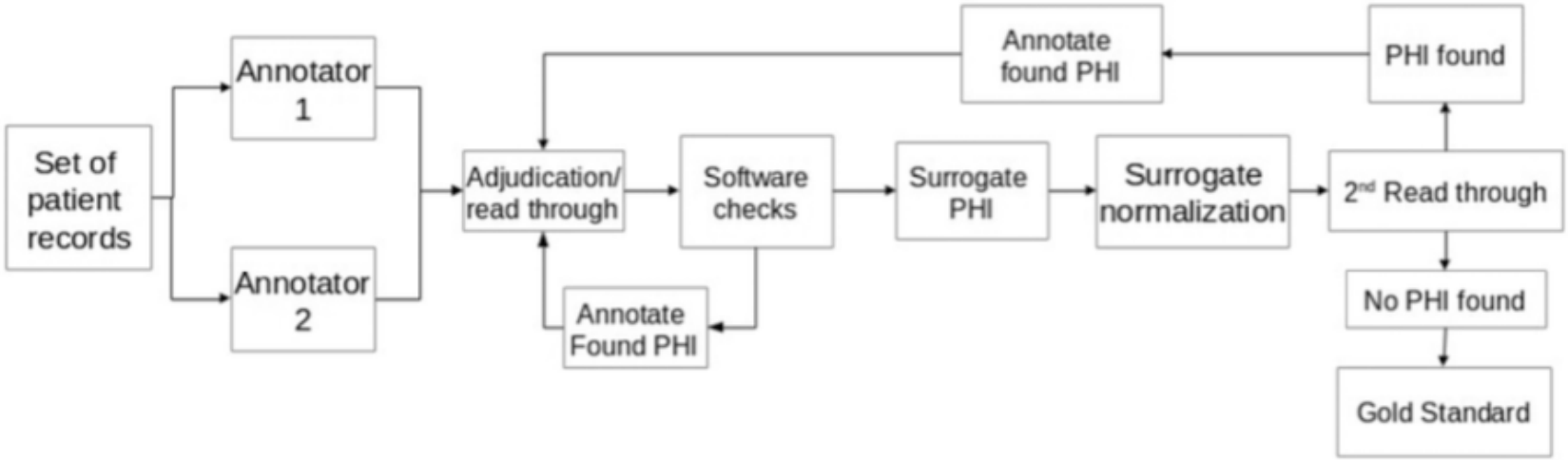
Table 2
Comparison of token counts in 2016 and 2014 shared task corpora.

	2016	2014
Total tokens	1,862,452	805,118
Average per record	1,862.4	617.4
Max	4610	2984
Min	304	617

Table 3
PHI category distributions between 2016 training and testing data, and comparison of PHI category totals between 2016 and 2014 corpora.

PHI category	Total #: 2016	Total #: 2014
NAME: PATIENT	2107	2195
NAME: DOCTOR	3963	4797
NAME: USERNAME	25	356
PROFESSION	2481	413
LOCATION: HOSPITAL	3523	2312
LOCATION: ORGANIZATION	1810	206
LOCATION: STREET	80	352
LOCATION: CITY	2214	654
LOCATION: STATE	1143	504
LOCATION: COUNTRY	1042	183
LOCATION: ZIP CODE	40	352
LOCATION: OTHER	44	17
AGE	5991	1997
DATE	9544	12,487
CONTACT: PHONE	256	524
CONTACT: FAX	9	10
CONTACT: EMAIL	7	5
CONTACT: URL	8	2
CONTACT: IPADDRESS	0	0
ID: SSN	0	0
ID: MEDICAL RECORD	6	1033
ID: HEALTH PLAN	2	1
ID: ACCOUNT	0	0
ID: LICENSE	59	0
ID: VEHICLE	0	0
ID: DEVICE	0	15
ID: BIO ID	0	1
ID: ID NUMBER	10	456
Total # of PHI phrases	**34,364**	**28,872**
Average PHI per file	**34**	**22.14**

The DEID Annotation Pipeline from i2b2 2014 (Stubbs et al, 2014).



Design Considerations for DEID Studies

- How many patients and notes to sample? Generalizability of sample?
- What definition of PHI, and how to define labels in annotation process.
- Evaluating the annotation process: inter/intra-rater agreement, time/cost, etc.
- Choice of algorithmic method for DEID: rule, ML, hybrid, DL, etc.
- Evaluation metrics for DEID: instance/token-level, BIO, SENS/PPV/F1.
- Train/val/test vs. Cross-val; Temporal/Geographic Hold-Outs; Generalizabily.
- Post-hoc error analysis. Are errors poor training data vs. poor model?
- Other issues: pre-annotation, PHI enrichment, over-scrubbing, etc.

Token-Level Evaluation Metrics for DEID

- Sensitivity (Recall): TP/(TP+FN)
- Positive Predictive Value (Precision): TP/(TP+FP)
- F1 (harmonic mean recall/precision): (2*P*R)/(P+R)

Instance-Level Evaluation Metrics for DEID

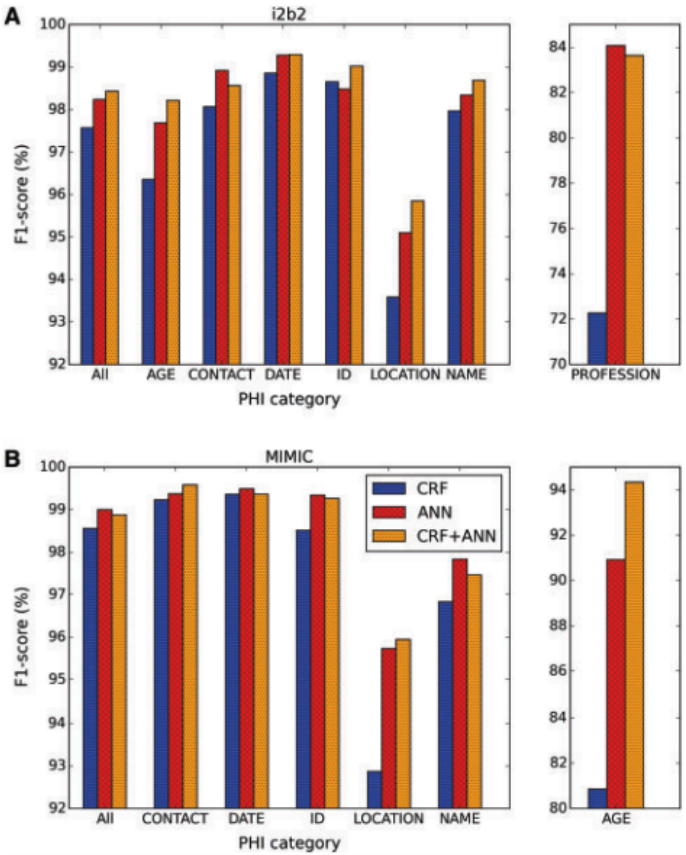
- 3 error types: (1) substitution, (2) insertion, and (3) deletion.
- Instance-level PHI has 3 slots: (1) type, (2), content and (3) extent.
- 4 values per slot: (1) correct, (2) incorrect, (3) missing, and (4) spurious.
- Let {C,I,D,S} represent correct and insertion/deletion/substitution errors, respectively.
- Instance level precision (P): C/(C+S+I)
- Instance level recall (R): C/(C+S+D)
- Instance level F1: (2*P*R)/(P+R)

Aggregation of Evaluation Metrics

- Micro-averaged. Estimate evaluation metrics for entire corpus.
- Macro-averaged. Estimate evaluation metrics for each document, and average.

Error Analysis for DEID Systems

- Errors by PHI type.
- Location/organization/profession hard.
- Name mid-easy.
- PHI with simple patterns easy.



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 on the test set. We discuss the error categorization scheme in Section 5.2.1 and present the results in Section 5.2.2.

5.2.1 Error Categorization. We distinguish between two 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. [5], we consider the following sources of modeling errors:

- **Abbreviation.** PHI instances which are abbreviations or 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 a given 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 or exclude those prefixes.
- **Common Language.** PHI instances consisting of common language 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 is 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.

Systematic Review (Meystre, 2010): State of the Art is a Hybrid ML System.

Table 4 Algorithms and features used by systems mostly based on machine learning methods

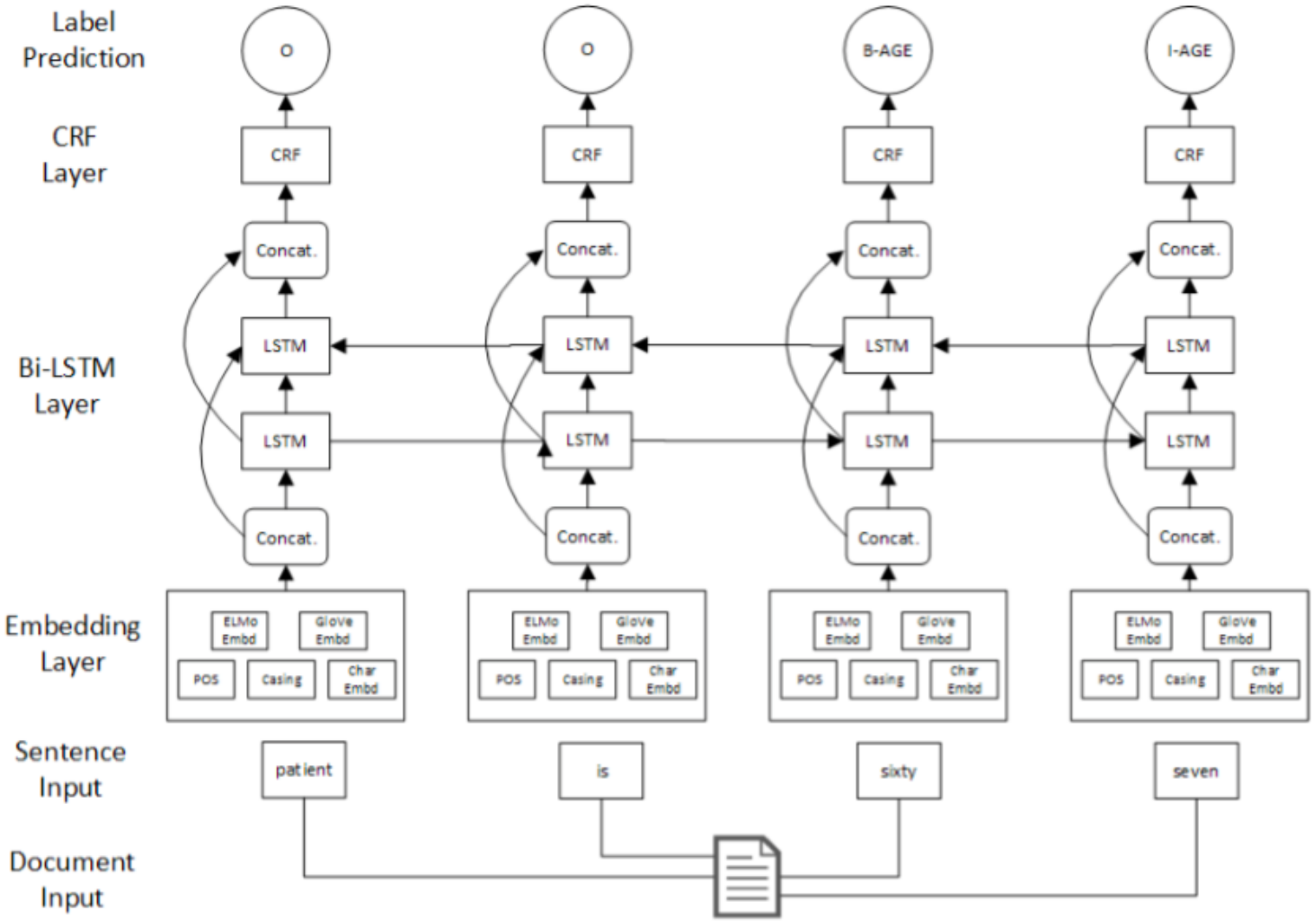
De-identification system	Machine learning algorithm	Features		
		Lexical/morphological	**Syntactic**	**Semantic**
Aramaki	CRF	Word, surrounding words (5 words window), capitalization, word length, regular expressions (date, phone), sentence position and length.	POS (word + 2 surrounding words)	Dictionary terms (names, locations)
Gardner	CRF	Word lemma, capitalization, numbers, prefixes/suffixes, 2-3 character n-grams	POS (word)	None
Guo	SVM	Word, capitalization, prefixes/suffixes, word length, numbers, regular expressions (date, ID, phone, age)	POS (word)	Entities extracted by ANNIE (doctors, hospitals, locations)
Hara	SVM	Word, lemma, capitalization, regular expressions (phone, date, ID)	POS (word)	Section headings
Szarvas	Decision Tree	Word length, capitalization, numbers, regular expressions (age, date, ID, phone), token frequency	None	Dictionary terms (first names, US locations, countries, cities, diseases, non-PHI terms), section heading.
Taira	Maximum Entropy	Capitalization, punctuation, numbers, regular expressions (prefixes, physician and hospital name, syndrome/ disease/procedure)	POS (word)	Semantic lexicon, dictionary terms (proper names, prefixes, drugs, devices), semantic selectional restrictions
Uzuner	SVM	Word, lexical bigrams, capitalization, punctuation, numbers, word length.	POS (word + 2 surrounding words), syntactic bigrams (link grammar)	MeSH ID, dictionary terms (names, US and world locations, hospital names), section headers.
Wellner	CRF	Word unigrams/bigrams, surrounding words (3 words window), prefixes/suffixes, capitalization, numbers, regular expressions (phone, ID, zip, date, locations/ hospitals)	None	Dictionary terms (US states, months, general English terms).

CEGS-NGRID (Stubbs, 2016): State of the Art is a Deep Learning Model

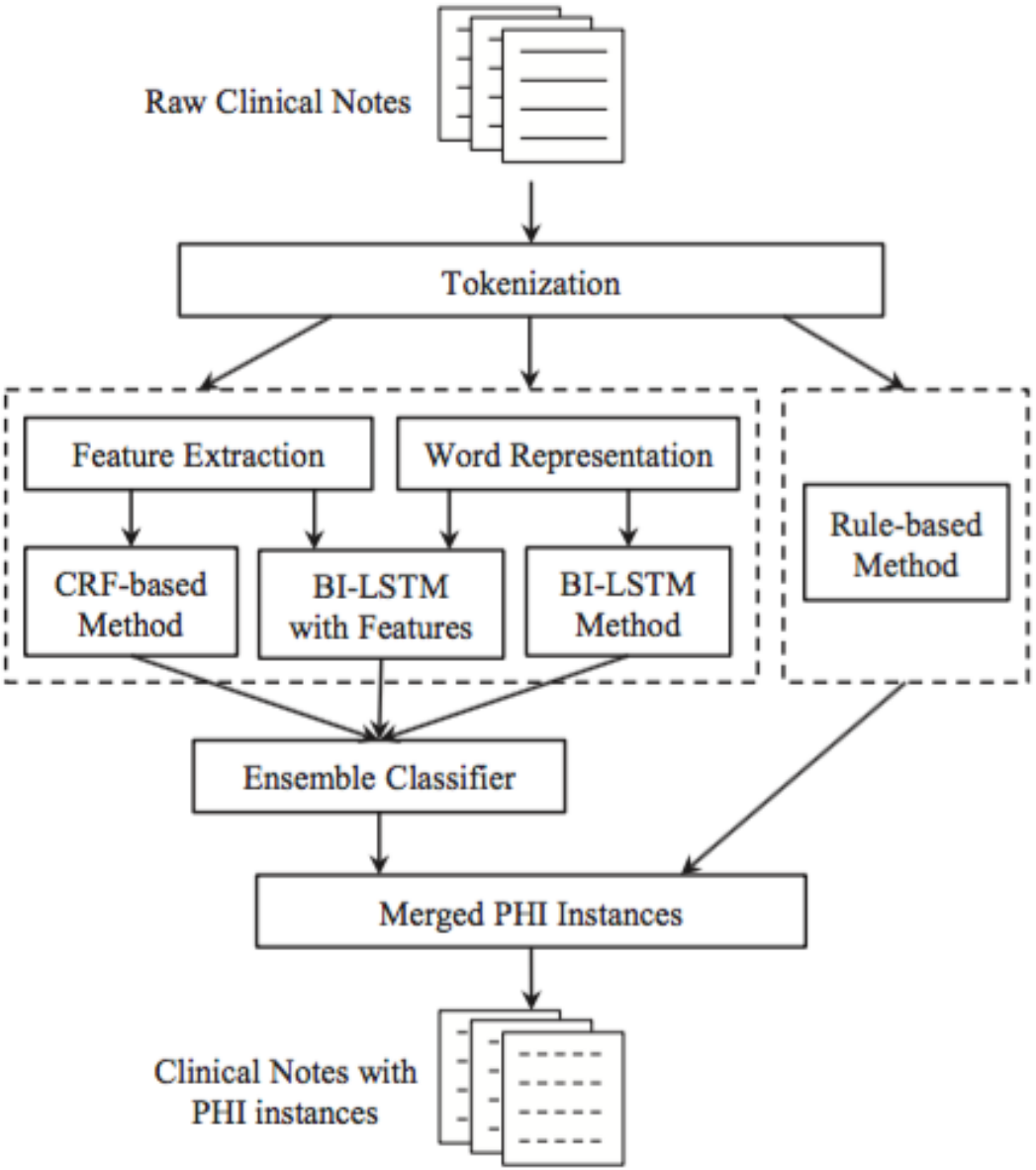
Rank	Team	Systems and processing	Precision	Recall	F1
1	Harbin Institute of Technology Shenzhen Graduate School	**Pre-processing**: Tokenization **System**: 4 modules: – CRF – 2 BI-LSTM – Rules **Post-processing**: merged CRF and BI-LSTM outputs with SVN ensemble classifier, then incorporated rules [20]	0.9422	0.8881	0.91430
2	University of Texas Health Science Center at Houston	**Pre-processing**: tokenization, POS tagging, section parsing **System**: – Rules – 2 CRFs: one for numbers, one for names **Post-processing**: merging output, error correction [16]	0.9339	0.8823	0.90740
3	The University of Manchester	**System**: Combined outputs from two systems: – mDEID, an CRF-based system [5] – CliDEID, a 'data-driven' CRF system [8] **Post-processing**: combined outputs of two systems; kept longer of overlapping spans [6]	0.8888	0.8653	0.87690
4	University of Alabama at Birmingham	**Preprocessing**: tokenization, sentence and section detection **System**: multi-pass "sieve" system: – pattern matching – dictionary matching – Stanford CRF [10,3]	0.9162	0.8338	0.87310
5	Harbin Institute of Technology	**Pre-processing**: tokenization, sentence detection **System**: BI-LSTMs developed tags for each token **Post-processing**: a CRF layer identified most likely tag for each token [31]	0.8418	0.8728	0.85700
6	National Taitung University	**Pre-processing**: tokenization **System**: CRF [13]	0.7958	0.8501	0.82210
7	MITRE	System: MIST [4] with additional lexicons [2]	0.8552	0.762	0.80590
8	LIMSI-CNRS	**Pre-processing**: two text segmentations, token- and character-based **System**: 2 CRFs, one for each text segmentation **Post-processing**: merged outputs, prioritized character-based [12]	0.847	0.6963	0.76430
9	National Central University	**Pre-processing**: sentence detection, tokenization, POS tagging and chunking **System**: combination of rules and CRF **Post-processing**: dictionary matching, rules [25]	0.7892	0.779	0.75960
10	University of Utah #77	**Pre-processing**: sentence detection, tokenization, POS tagging, chunking **System**: Pipeline of dictionary matching, rules, and CRF [1]	0.8645	0.6132	0.71750

biLSTM-CRF Architecture (Khin, 2018)

Figure 1: Deep Learning Architecture



biLSTM-CRF Architecture (Liu, 2017)



Systematic Review (Yogarajan; 2018, 2020): Open Research Questions

5.3. Research Questions

Although these de-identification competitions have provided a great platform to develop this area, some questions/areas are worth exploring.

(1) Considering the noticeable changes across the systems - is the hybrid system the best way to go? Could machine learning based system (for example, LSTM) outperform and be more adaptable to a new dataset?
(2) Is the use of rules and regular expressions crucial for rare, occurring PHIs? If so can we learn rules and regular expressions from data instead of handwritten rules (as seen in these competitions)?
(3) Apart from the noted fact that some PHIs, such as PROFESSION and LOCATION, perform poorly due to the low occurrence in training data, is there any other reason for this? Especially considering the use of handwritten rules and dictionary in the competitions, what are some other possible issues related to the poor F-measure? Any possible solutions to these problems?
(4) Adaptability to a new dataset is a noticeable issue with the de-identification systems. This is partly due to the lack of datasets. Could transfer learning be used to solve this problem? If so, is there any similar datasets available from domains such as law and crime science?
(5) Is there a way to obtain similar results to that of 2016 competitions with less complicated, reproducible systems?
(6) Considering the 95% on HIPAA regulations is still to be achieved across all PHIs - is there a need for stricter regulations, such as that introduced in these competitions? If so the re-identification implications will be an interesting avenue to explore.
(7) Systems developed using 2014 datasets after the competitions reported better performances for poorly performing PHIs. Is there a reason for this change? Would these systems do equally well if used for a different dataset? Can techniques from these systems be adopted to develop a more robust and better performing de-identification system?
(8) Would approaches such as word embeddings, which try and model the semantics of words, better be able to capture rarely occurring PHIs as opposed to syntactic approaches (given the lack of examples in the datasets to model the syntax of the rare PHIs)?

A Selection of Open Source DEID Tools

- ICES/UManchester. MASK
- PhysioNet/MIT: deid
- Carnegie Mellon: deidentify
- MITRE Corp: MIST

References

Dernocourt, F., Lee, J., Uzuner, O., Szolovits, P. (2007). De-Identification of Patient Notes with Recurrent Neural Networks. JAMIA.

Meystre, S., Friedlin, J., South, B., Shen, S., Sarmore, M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology.

Neahmatullah, I., Douglass, M., Lehman, L., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R., Clifford, G. (2008). Automated De-Identification of Free Text Medical Records. BMC Medical Informatics and Decision Making.

Stubbs, A., Uzuner, O. (2014). Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth Corpus. Journal of Biomedical Informatics.

Stubbs, A., Filannino, M., Uzuner, O. (2016). De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics.

Uzuner, O., Luo, Y., Szolovits, P. (2007). Evaluating the State of the Art in Automatic De-Identification. JAMIA.

Yogarajan, V., Mayo, M., Pfahringer, B. (2018). A Survey of Automatic De-Identification of Longitudinal Clinical Narratives. Arxiv.