Journal of Biomedical Informatics 58 (2015) S11–S19

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1

Amber Stubbs [a,*], Christopher Kotfila [b], Özlem Uzuner [b]

[a] School of Library and Information Science, Simmons College, Boston, MA, USA
[b] Department of Information Studies, State University of New York at Albany, Albany, NY, USA

ARTICLE INFO

Article history:
Received 10 March 2015
Revised 3 June 2015
Accepted 4 June 2015
Available online 28 July 2015

Keywords:
Natural language processing
Machine learning
Shared task
Medical records

ABSTRACT

The 2014 i2b2/UTHealth Natural Language Processing (NLP) shared task featured four tracks. The first of these was the de-identification track focused on identifying protected health information (PHI) in longitudinal clinical narratives. The longitudinal nature of clinical narratives calls particular attention to details of information that, while benign on their own in separate records, can lead to identification of patients in combination in longitudinal records. Accordingly, the 2014 de-identification track addressed a broader set of entities and PHI than covered by the Health Insurance Portability and Accountability Act – the focus of the de-identification shared task that was organized in 2006. Ten teams tackled the 2014 de-identification task and submitted 22 system outputs for evaluation. Each team was evaluated on their best performing system output. Three of the 10 systems achieved F_1 scores over .90, and seven of the top 10 scored over .75. The most successful systems combined conditional random fields and hand-written rules. Our findings indicate that automated systems can be very effective for this task, but that de-identification is not yet a solved problem.

1. Introduction

The 2014 i2b2[1]/UTHealth[2] Natural Language Processing (NLP) shared task featured four tracks. The first of these was the de-identification track focused on identifying protected health information (PHI) in the clinical narratives. While identifying PHI for removal, it is important for de-identification to preserve the medically salient contents of the narratives so that this information can benefit downstream research and maintain the value of the record for the care of the patients.

The 2014 shared task data were selected to show the progression (or lack thereof) of heart disease in diabetic patients over time, the focus of Track 2 of the i2b2/UTHealth shared task [1]. In order to reflect the progression over time, the records were longitudinal: the same patients were represented over multiple documents separated by weeks, months, or years. The inclusion of longitudinal records in a corpus presents a unique challenge for de-identification: Including more records from a patient's medical record provides important medical data for clinical research, but it also potentially puts the patient at greater risk of being identified.

America's Health Insurance Portability Accountability Act (HIPAA; 45 CFR 164.514) defines 18 categories of PHI, which must be removed from a medical record before it can be considered safely de-identified. These categories include patient names, contact information, ID numbers, and so on. However, a recent study in Canada showed that over an 11-year period, records of people's addresses alone could lead to their being identified [2]. Similarly, US citizens can be identified by their date of birth, ZIP code, and gender [3,4], yet the HIPAA PHI categories do not include gender, years, or full ZIP codes for sufficiently populated areas. In other words, while HIPAA provides a starting point for effective de-identification, it may not be sufficient for full de-identification.

While full de-identification may not be a realistic and attainable goal, expanding HIPAA categories to include a wider set of information can make de-identification more secure. Accordingly, the 2014 i2b2/UTHealth shared task data were de-identified to a more strict standard than what HIPAA defines [5,6] using additional categories for PHI, such as professions, full dates, and information about medical workers and facilities. We refer to this expanded set of PHI categories as i2b2-PHI categories (see Section 3).

We defined the Track 1 shared task consistently with the de-identification that we performed for data release. We released

* Corresponding author at: School of Library and Information Science, Simmons College, 300 The Fenway, Boston, MA 02115, USA. Tel.: +1 617 521 2807.
E-mail address: stubbs@simmons.edu (A. Stubbs).
[1] Informatics for Integrating Biology and the Bedside.
[2] University of Texas Health Science Center at Houston.

http://dx.doi.org/10.1016/j.jbi.2015.06.007
1532-0464/

60% of the de-identified data, with the gold standard i2b2-PHI annotations (but after the authentic PHI were replaced with realistic surrogates) as the training corpus. We gave the participants three months to build systems that automated the de-identification task. At this point, we released the remaining data, without annotations, as test data, and gave the participants three days to submit up to three system runs on the test data. We evaluated the system runs on two sets of PHI categories: the 18 categories defined by HIPAA (HIPAA-PHI) and the i2b2-PHI. We ranked the systems primarily based on their performance on the i2b2-PHI.

This paper provides a brief overview of the de-identification task (Track 1) of the i2b2/UTHealth 2014 shared task, related work (Section 2), data (Section 3) and annotation (Section 4). Its focus is primarily on the evaluation metrics (Section 5), descriptions of participating systems (Section 6) and results of the shared task. To put this task into context, we compare these results to the results of the 2006 i2b2 de-identification task (Section 7) and close the paper with a discussion and conclusions (Sections 8 and 9).

2. Related work

There have been many shared tasks in NLP, but few are comparable to the 2014 de-identification task described here. Traditional named entity recognition (NER) is similar to de-identification, as the focus for both tasks is to identify information such as names, dates, and locations in texts. However, de-identification of medical records includes more categories of information than traditional NER, such as phone numbers, ID numbers, and ages. The 6th and 7th Message Understanding Conferences (MUCs) included shared tasks in NER. Specifically, the participants were asked to label entities (organizations, persons, locations), numbers (currencies and percentages), and temporal expressions (specific dates and times) [7,8]. The MUC-6 NER task participants included 20 systems from 15 teams, and 96.42 as the highest *F*-measure [7]. MUC-7 had 14 systems from 12 teams, with a top *F*-measure of 93.39 [9]. However, both of the MUC tasks were run on newswire texts, rather than clinical notes, making a direct comparison to the 2014 i2b2/UTHealth de-identification challenge untenable.

NER-type shared tasks in the biomedical domain tend to focus on identifying information related to the field, rather than traditional named entities. BioCreAtIvE [10] participants identified and mapped gene and protein names, and the TREC Genomics tracks also focused on genes and diseases when looking for particular entities [11]. The BioNLP'09 [12] focused on protein and gene event extraction, which the BioNLP'11 [13] and BioNLP'13 [14] tasks expanded upon. Each of these shared tasks used text from MEDLINE. Other biomedical shared tasks include the BioASQ tasks, which use data from PubMed [15].

To the best of our knowledge, the only other de-identification shared task made open to the public is the previous i2b2 event, held in 2006 [16]. The 2006 task used 889 de-identified records, one record per patient, and fielded sixteen submissions from seven teams. The 2006 data used individual records for each patient. As we noted in the Introduction, longitudinal records may contain much more personal information about a patient than individual records. And this information, while perfectly HIPAA-compliant and ineffective for identifying the patient when found in individual records and on their own, can be used collectively to piece together the identity of the patient over several records. This makes de-identification of longitudinal records a potentially more intricate task.

2.1. De-identification tools

In order to provide context for comparing the i2b2 participants with other recently developed de-identification systems, here we discuss three recent systems and their results. A broader overview of de-id systems can be found in the recent review article by Meystre et al. [17], in which the authors describe 18 de-identification systems built between 1995 and 2010. Here, we focus on three recent tools: MIST, the MITRE Identification Scrubber Toolkit [18], BoB, the "best of breed" tool from the Veteran's Health Administration [19], and an in-house tool from Cincinnati Children's Hospital Medical Center [20].

MITRE's MIST tool [18] is an open source de-identification system that also includes annotation and PHI replacement tools. The parts of the system that identify PHI use the Carafe engine [21], a system that uses a Conditional Random Field (CRF) [22] model trained specifically for text processing. The Carafe engine is the only system used in MIST: it does not implement rules, though in the conclusions the authors note that some types of PHI may be better captured through rules. When run on the 2006 i2b2 data, MIST achieved precision of 0.978, recall of 0.951, and F_1 of 0.965.

The VHA's BoB [19] is built on the Apache UIMA architecture [23] and uses cTAKES [24] to pre-process the documents. The system then uses a "stepwise hybrid" approach to removing PHI. In the first step, a "high sensitivity extraction component", uses rules and a CRF model to identify all possible PHI in a document. In the second step, a "false positives filtering component" uses Support Vector Machine (SVM) [25] classifiers to remove inaccurate PHI tags generated in the first step. When tested against the 2006 i2b2 corpus, and implementing special rules to account for the differences in annotations, BoB achieved precision, recall, and F_1 of 0.846, 0.965, and 0.902, respectively.

The Cincinnati Children's Hospital Medical Center's (CCHMC) in-house de-identification system [26] is based on the MALLET package [27], which also uses CRF models. The CCHMC system also utilizes pre-processing in the form of an in-house and the TreeTagger[3] part of speech processor, and post-processing in the form of rules that identify email addresses, match names to an external lexicon, and capture any names that the CRF module missed. When tested on the 2006 i2b2 corpus, with training data from other corpora, the system achieved precision, recall, and F_1 of 0.9682, 0.9342, and 0.9509, respectively [20].

Overall, these systems perform quite well, and set a high standard for further research in de-identification. Many differences in the scores can be attributed to differences in training data, as each group had access to data that was unavailable to the others at the time.

3. Data

The data for this task are a newly de-identified corpus of longitudinal medical records, drawn from the Research Patient Data Repository of Partners Healthcare [28]. This corpus was used for all the tracks of the 2014 i2b2/UTHealth shared task. It consists of 1,304 medical records for 296 diabetic patients. All PHI in these records have been removed [5,29] and replaced with realistic surrogates [6]. This shared task was open to all interested researchers from any country, and was announced on various mailing lists in the NLP community, as well as a mailing list of past i2b2 shared task participants. We released

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html.

approximately two thirds of the training data in May 2014, and released the remaining third in June. In early July we released the test data, at which time participants were asked to stop developing their systems, and they were given three days to submit up to three system runs on the test data. The corpus was distributed to the shared task participants under a data use agreement and will be available to the rest of the community for research from https://www.i2b2.org/NLP/ in November 2015. Institutional review boards at Partners Healthcare, MIT, and SUNY Albany approved this study.

4. Annotation

As we described in the Introduction, due to the longitudinal nature of our data, we were aware that small amounts of information about the patients that would not be considered PHI under HIPAA could be pieced together to reveal a person's identity. Therefore, to ensure the patients' protection as much as possible, we used HIPAA-PHI categories as our starting point, augmented and added sub-categories, and created the following i2b2-PHI categories with their "type" attributes:

- NAME (types: PATIENT, DOCTOR, USERNAME)
- PROFESSION
- LOCATION (types: ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER)
- AGE
- DATE
- CONTACT (types: PHONE, FAX, EMAIL, URL, IPADDRESS)
- IDs (types: SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER)

Of these i2b2-PHI categories, only the following correspond to the HIPAA-PHI categories: NAME-PATIENT, LOCATION-STREET, LOCATION-CITY, LOCATION-ZIP, LOCATION-ORANIZATION, AGE, DATE, all ID sub-categories as well as CONTACT-PHONE, CONTACT-FAX, CONTACT-EMAIL.

Given these PHI categories and types, we annotated the information in each record twice, and implemented a series of automatic and manual checks to ensure that all authentic PHI were annotated. We replaced all annotated authentic PHI with realistic surrogates, and re-checked the records for readability [5]. We used these annotated data as the source of the training and testing data released to the participants, and as the gold standard against which we evaluated the system outputs, as we describe in the next section.

5. Evaluation

We used precision (Eq. (1)), recall (Eq. (2)) and *F*-measure (Eq. (3)) scores to evaluate the participants' results against the gold standard annotations. We checked the significance of the differences of the systems from each other using approximate randomization [30,31].

$$\text{Precision (P)} = \text{true positives}/(\text{true positives} + \text{false positives}) \quad (1)$$

$$\text{Recall (R)} = \text{true positives}/(\text{true positives} + \text{false negatives}) \quad (2)$$

$$F\text{-measure } (F_1) = 2 * ((P * R)/(P + R)) \quad (3)$$

We calculated P, R, and F_1 at both entity and token levels across the entire corpus. We used micro-averaged F_1 as our primary metric. The evaluation scripts we used are freely available on GitHub: https://github.com/kotfic/i2b2_evaluation_scripts/tree/v1.2.1.

Entity-based (also known as "instanced-based" [16]) evaluations require that system outputs match the beginning and end locations of each PHI tag exactly, as well as match the tag name and type attribute. Token-based evaluations must also match the tag's name and type attribute, but are evaluated on a per-token basis. In other words, if the gold standard has "Rayna De Angelis" annotated as NAME with type DOCTOR, and a system annotated "Rayna" and "De Angelis" as individual tags that each have NAME/DOCTOR annotations, the entity-based evaluation would not count that system's output as correct, though the token-based evaluation would. We perform both entity- and token-based evaluations because entity-based evaluations are the standard system for named entity recognition, where it is important for a phrase describing an entity to be captured whole. However, for the purposes of de-identification, it is less important for all parts of a single PHI to be identified together as long as all the parts are identified as PHI at some point. As long as "Rayna" and "De Angelis" are identified as PHI and removed from the corpus, it does not matter if that is the result of a single entity annotation or two token-based annotations.

Given these metrics, we take into consideration the differences in de-identification requirements of medical institutions for data release for our evaluations. For example, some institutions may want any possible PHI removed, while others may be only concerned with those categories specified by HIPAA. Therefore, we evaluated systems on both the i2b2-PHI and the HIPAA-PHI.

While we are concerned with the correct recognition of specific PHI categories from the perspective of preserving the integrity of the data, even in the absence of correct identification of categories of PHI, de-identification can succeed. In other words, systems can effectively remove PHI and preserve patient privacy without correctly differentiating between PHI categories. For example, a system that incorrectly identifies patient names as doctor names will still successfully remove the identified names from the records and accomplish de-identification even if it gets to that correct outcome for the wrong reasons. In order to evaluate systems purely on PHI identification without categories, we performed a binary evaluation on the recognition of PHI vs non-PHI (binary PHI categories).

As a result, we evaluated performance in the following ways: (1) micro-averaged entity- and token-based P, R, F_1 on i2b2-PHI categories. This evaluation determines how well each system did compared to the gold standard. (2) Micro-averaged entity- and token-based P, R, F_1 on only the HIPAA-PHI categories. We perform this evaluation to determine whether a system's performance is good enough for meeting HIPAA requirements. (3) Micro-averaged entity- and token-based evaluation of binary PHI categories. We perform this evaluation to check whether the records are de-identified effectively, even if for the wrong reasons. We used the micro-averaged entity-based F_1 over i2b2-PHI categories as our primary ranking metric.

Table A2 in the Appendix A shows, for a given ground truth, whether it would be considered correct (marked with a +) or incorrect (marked with a −) under each of our evaluation metrics. As can be seen, the token-based evaluations are the most accepting of variations in the system outputs (marked with the highest number of +s), while entity-based analyses require much stricter adherence to the gold standard in terms of matching the starting and ending offsets of every tag exactly. The most difficult task is the entity-based evaluation over the i2b2-PHI. In contrast, the binary evaluations accept any annotation that identified PHI.

We calculated statistical significance between system runs using approximate randomization as outlined in Chinchor [30]

and Noreen [31]. Significance was tested for micro-averaged P, R and F_1, with $N = 9999$ and an alpha of 0.1. These values are consistent with MUC-3 and MUC-4 evaluations as well as previous i2b2 challenges [16].

6. Submissions

Each participating team submitted up to three system outputs for evaluation to the de-identification track of the 2014 i2b2/UTHealth shared task. Overall, we received 22 submissions from 10 teams (see Table A1 in the Appendix A for details on participating teams, their members and affiliations). The most popular and successful approaches among the submissions were hybrids of Conditional Random Fields (CRFs) and hand-written rules, which processed the outputs of the two different systems into a coherent whole. We present the system overviews here alphabetically by team name. Two of the teams (East China Normal University and UC San Diego) did not submit system descriptions and are accordingly omitted from this overview.

The team from Harbin Institute of Technology [32] pre-processed their data with the OpenNLP[4] system's sentence detector and tokenizer, along with some regular expressions to tokenize irregular phrases. They then trained a CRF system on the following features: lexical, orthographic, and syntactic. Unlike most other systems, they did not use any medical dictionaries to identify key words.

The Harbin Institute of Technology Shenzhen Graduate School [33] team used three systems to generate annotations. First, a CRF based on token-level features, which used MedEx [34] for tokenization, and included features such as bag-of-words, part of speech, orthographic features, section information, and word representation features. Second, a CRF based on character-level features to extract PHI represented by characters, which used similar features to the token-level classifier, but decomposed raw sentences into characters instead of tokens. Third, a rule-based system that used regular expressions to identify standardized PHI such as PHONE, FAX, and MEDICAL RECORD NUMBER. They used a rule-based system to merge the outputs of the three systems: non-overlapping PHI instances were included directly in the system output; overlapping output from the three systems was resolved in a hierarchy, with preference given first to the rule-based classifier, then the character-level classifier, then the token-level classifier.

The team from Kaiser Permanente (Torii et al. [35]) focused on adapting the MIST tool[5] for the 2014 shared task data. They added their own annotated data to the 2014 shared task training medical records and augmented the MIST tool by providing additional rules that used lexicons (for LOCATION and PROFESSION categories) and regular expressions (for PHONE, ZIP, and ORGANIZATION categories). These rules also prevented certain types of non-PHI, such as font names, from being annotated. In addition to MIST, the team also trained a NER model on the 2014 shared task data using the Stanford NER system.[6] Their best run then merged the outputs of MIST and Stanford NER systems by taking the longest span of overlapping outputs from them.

The LIMSI–CNRS team (Grounin [36]) trained a CRF with different linguistic categories of features. Surface features represented information such as the token itself, token length, typographic case, presence of punctuation or digits. For morpho-syntactic features, they used part of speech categories obtained from Tree Tagger.[7] They identified semantic types by using trigger words from different categories (e.g., Dr., MD, Mr., Mrs., etc), as well as a list of professions from Wikipedia. They also used distributional analysis features, such as frequency in the corpus, document section, and cluster ID based on context. They then used 77 regular expressions to correct CRF outputs by, for example, identifying multi-word expressions and multi-token sequences by comparing them to a lexicon collected from the training corpus and fixing annotation spans for AGE and DOCTOR tags. They submitted three system runs: CRF only, CRF + rules without the lexicon from the training corpus, and CRF + rules with the lexicon. Their best run used the CRF + rules with the lexicon.

The UNIMAN team from Manchester [37] pre-processed the input data with cTAKES[8] and GATE[9] for tokenization, sentence splitting, part-of-speech tagging, and chunking. They built a combined knowledge- and data-driven system for identifying PHI. The knowledge-based component used dictionaries and a small set of rules (with orthographic, pattern, negation, lexical and context features, e.g., words from specialized vocabularies, symbols, and special characters). The data-driven component used a CRF model for each of the following categories: CITY, DATE, HOSPITAL, ORGANIZATION, PROFESSION, and PATIENT. Their CRF features included lexical (lemma, part of speech for the token and surrounding words), orthographic (capitalization, digits; orthographic patterns), semantic (matched to the dictionaries of related vocabulary), and positional features (position in line, presence of space between current and next token). They also proposed a two-pass approach for some categories (PATIENT, DOCTOR, HOSPITAL, CITY, MEDICAL RECORD, and ID_NUM): for each category, they extracted the initial annotations at the patient-level and created a run-time patient-specific dictionary. This dictionary was subsequently used for 'second-pass' dictionary matching on the same set of patient narratives in order to capture mentions not recognized in the initial pass. Finally, an integration step merged the outputs of the rule-based and CRF modules using different sets of rules for the different system runs. The most successful merging system used rules for DATEs and DOCTORs, and rules and lexicons for PATIENTs.

The Newfoundland team (Chen [38]) used a non-parametric Bayesian [39] Hidden Markov Model (HMM) [40]. This model utilizes latent variables to organize words of the same label into more refined categories, which allows the model to capture subtle variations in the data. Instead of using a fixed number of latent variables, which makes a strong assumption of data, the model allows an infinite number of latent variables by implementing a Dirichlet process [41] as a prior and lets the data determine the optimum number of latent variables. The Newfoundland system implemented a Dirichlet process to identify PHI. They also implemented a set of features to identify words that did not appear in the training data.

The team from Nottingham (Yang [42]) pre-processed the data via sentence splitting, tokenization, part of speech tagging, and shallow parsing. They then identified the following features: word-token (lemma, part of speech (POS), chunk), context (lemma, POS, chunk of nearby tokens), orthography (capitalization, punctuation, regex patterns for dates, usernames, etc.), sentence-level features (position of token in sentence, section headers), task-specific features (lists of names and acronyms of US states, countries, languages, and lexical clues such as presence of "Dr." or "M.D."). Their system is a hybrid one: they trained a CRF using the described features, and then used dictionaries and regular expressions to identify PHI with few sample instances. As a post-processing step, they performed entity extraction from identified PHI, and used a trusted PHI term list to uncover more potential terms. They generated the

[4] https://opennlp.apache.org/.
[5] http://mist-deid.sourceforge.net/.
[6] http://nlp.stanford.edu/software/CRF-NER.shtml.
[7] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html.
[8] http://ctakes.apache.org/.
[9] https://gate.ac.uk/.

trusted PHI term list by making use of different types of relations between detected PHI terms.

The San Marcos (Guillen [43]) team built an entirely rule-based system. They performed an analysis of the data to determine the most frequently used tags, tokenized the texts, performed rule-based token classification based on whether the tokens were digits or non-digits, and analyzed the result to determine patterns for DATEs, AGEs, and IDs. They refined their rules with lexical clues.

Table 1 shows an overview of the different tools, rules, machine learning algorithms, features, and external resources used by these systems.

7. Results

We evaluated all systems on both i2b2-PHI and HIPAA-PHI categories, both based on entity- and token-level annotations. We ranked each team based on their top performing system run in terms of micro-averaged entity-based F_1 on i2b2-PHI categories. The evaluations reported in this section are the results from the shared-task submissions; we report results only on the best-performing submission of each team. In their own papers in this supplement, the participants had the opportunity to present systems and results that improve upon their shared task submissions.

Table 1
Overview of tools, rules, machine learning, and external resources used in systems.

Team name	Tools	Rules and features	Machine learning systems and features	External resources
Harbin Institute of Technology	OpenNLP CRF++	Regular expressions for tokenization	CRF: lexical, syntactic, orthographic	
Harbin Institute of Technology Shenzhen Graduate School	MedEx	Regular expressions for categories such as PHONE, FAX, MEDICAL RECORD, EMAIL and IPADDR	CRFs: bag-of-words; part-of-speech (POS) tags; combinations of tokens and POS tags; sentence information; affixes; orthographical features; word shapes; section information; dictionary features	
Kaiser Permanente	MIST Stanford NER	Regular expressions for categories such as PHONE, EMAIL, ZIP	MIST, Stanford NER; features not mentioned	Personal de-id corpus
LIMSI–CNRS	Tree Tagger MEDINA toolkit	Rules to correct output of CRF	CRF: surface features, morpo-syntactic, semantic, distributional	
UNIMAN	Pre-processing: CTAKES and GATE	JAPE system: orthographic, pattern, contextual, entity	CFR: lexical, orthographic, semantic, positional	Dictionaries collected from Wikipedia, GATE, and deid
Newfoundland	Python packages Numpy and Scipy		Non-parametric Bayesian Hidden Markov Model: token, word token, number token	
Nottingham	Pre-processing; CRF++	Yes, for categories such as FAX, EMAIL, DEVICE, BIOID	CRF: Word-token, context, orthographic, sentence-level, task-specific	Self-compiled dictionary
San Marcos		Used for all categories of PHI		

Table 2
i2b2-PHI categories: micro- and macro-averaged entity-based evaluation. The micro-F_1 scores are bolded, as they are the primary evaluation metric for this task.

Rank	Team name	Micro-averaged precision	Micro-averaged recall	Micro-averaged F_1	Macro-averaged precision	Macro-averaged recall	Macro-averaged F_1
1	University of Nottingham	0.9645	0.9092	**0.9360**	0.9653	0.9156	0.9398
2	Harbin Institute of Technology: Shenzhen Graduate School	0.9264	0.8988	**0.9124**	0.9282	0.9091	0.9185
3	University of Manchester	0.9306	0.8836	**0.9065**	0.9341	0.893	0.9131
4	Harbin Institute of Technology	0.9229	0.8505	**0.8852**	0.9228	0.8576	0.889
5	Kaiser Permanente	0.8731	0.7700	**0.8183**	0.8727	0.7627	0.814
6	LIMSI–CNRS	0.8937	0.7332	**0.8055**	0.8909	0.7272	0.8007
7	Memorial University of Newfoundland	0.7944	0.7191	**0.7549**	0.7963	0.7259	0.7595
8	East China Normal University	0.9344	0.4288	**0.5878**	0.9215	0.3951	0.5531
9	California State University San Marcos	0.7628	0.3526	**0.4823**	0.7217	0.3235	0.4467
10	UC San Diego	0.5271	0.3825	**0.4433**	0.5662	0.4011	0.4696

Table 3
Significance tests for de-identification systems using entity-based evaluations over i2b2-PHI categories.

	Nottingham	Harbin Grad	UNIMAN	Harbin	Kaiser	LIMSI–CNRS	Newfoundland	ECNU	San Marcos	UCSD
Nottingham										
Harbin-Grad	–									
UNIMAN	–	F_1								
Harbin	–	P	–							
Kaiser	–	–	–	–						
LIMSI–CNRS	–	–	–	–	–					
Newfoundland	–	–	–	–	–	F_1				
ECNU	–	–	P	–	–	–	–			
San Marcos	–	–	–	–	–	–	P	–		
UCSD	–	–	–	–	–	–	–	–	R	

Table 2 shows the entity-based results for each team's best system run on i2b2-PHI, sorted by micro-averaged F_1. Overall, the systems performed quite well on what is known to be a difficult task. Three systems achieved micro-averaged F_1 measures of over .9, and eight of them scored over .58. Given that the entity-based, i2b2-PHI evaluation has the strictest rules for obtaining a true positive, these results will always compare less favorably to other evaluations in this paper. We also include the macro-averaged scores in this table, which show that for most teams the macro evaluation scores barely differ from the micro scores for most teams. As this difference continues throughout the other evaluations, we omit macro scores from the rest of the paper.

Table 3 shows the results of the significance tests between the top-ranking submissions of each team, as determined by micro-averaged entity-based F_1 over i2b2-PHI categories (see Table 2). Cells with P, R, or F_1 indicate that the two systems are not significantly different in P, R, or F_1, respectively. Overall, we see that the majority of the systems are significantly different in terms of their output. Note that we only show half the table, as the upper diagonal would be symmetrically identical to the lower.

Table 4 shows the token-level evaluations of the i2b2-PHI categories. As we discussed, this is a less restrictive evaluation, as it counts each PHI token separately. All the teams' scores are higher using this evaluation metric, suggesting that some of the errors in the entity-based evaluations are from not capturing the entire PHI entity.

Table 4
i2b2-PHI categories: micro-averaged token-based evaluation.

Rank	Team name	Micro-averaged precision	Micro-averaged recall	Micro-averaged F_1
1	University of Nottingham	.9815	.9414	.9611
2	University of Manchester	.9722	.9250	.9480
3	Harbin Institute of Technology: Shenzhen Graduate School	.9564	.9366	.9464
4	Harbin Institute of Technology	.9571	.9051	.9304
5	Kaiser Permanente	.9397	.8609	.8986
6	LIMSI–CNRS	.9321	.7783	.8483
7	Memorial University of Newfoundland	.8629	.8038	.8323
8	East China Normal University	.9498	.5399	.6885
9	California State University San Marcos	.9010	.4753	.6223
10	UC San Diego	.7164	.4939	.5847

Table 5
HIPAA-PHI categories only: micro-averaged entity-based evaluation.

Rank	Team name	Micro-averaged precision	Micro-averaged recall	Micro-averaged F_1
1	University of Nottingham	.9763	.9390	.9573
2	Harbin Institute of Technology: Shenzhen Graduate School	.9513	.9307	.9409
3	University of Manchester	.9437	.9213	.9323
4	Harbin Institute of Technology	.9414	.8957	.9180
5	Kaiser Permanente	.8850	.8047	.8429
6	LIMSI–CNRS	.9137	.7666	.8337
7	Memorial University of Newfoundland	.8494	.7535	.7985
8	East China Normal University	.9335	.6117	.7391
9	California State University San Marcos	.7758	.4571	.5753
10	UC San Diego	.5384	.3900	.4524

Table 6
HIPAA-PHI categories only: micro-averaged token-based evaluation.

Rank	Team name	Micro-averaged precision	Micro-averaged recall	Micro-averaged F_1
1	University of Nottingham	0.9889	0.9629	0.9757
2	University of Manchester	0.9797	0.9542	0.9668
3	Harbin Institute of Technology Shenzhen Graduate School	0.9748	0.9578	0.9662
4	Harbin Institute of Technology	0.9708	0.9371	0.9536
5	Kaiser Permanente	0.9548	0.8914	0.9221
6	Memorial University of Newfoundland	0.9037	0.8365	0.8688
7	LIMSI–CNRS	0.9514	0.7972	0.8675
8	East China Normal University	0.9496	0.7273	0.8237
9	California State University San Marcos	0.9085	0.6039	0.7255
10	UC San Diego	0.7439	0.5089	0.6043

For comparison, Tables 5 and 6 show the results when the systems are evaluated on only the HIPAA-PHI categories at the entity and token levels, respectively. Table 4 shows that the token-based HIPAA-PHI evaluation resulted in the highest scores for each team. Again, as we would expect, the token-based evaluations result in higher scores, and overall the token-based HIPAA-PHI scores are the highest of all.

Fig. 1 shows the entity-based micro-averaged F_1 scores for the individual i2b2-PHI categories. Overall, the PROFESSION and LOCATION categories proved to be the most difficult. There are multiple factors that contribute to this. First, the phrases labeled as PROFESSION, LOCATION-ORGANIZATION and LOCATION-OTHER vary widely in content, form, and structure, from simple phrases such as "firefighter" or "Cape Cod" to complex descriptions such as "Ground Transit Operators Supervisor" or "Fountain Of The Four Rivers". Additionally, not all PROFESSIONs are nouns or noun phrases, making syntactic cues harder to use. For example, "nurse" and "nursing" are both labeled as PROFESSION because "she is a nurse" and "he works in nursing" both refer to a person's job. Lack of training data also contributes to the problem for these tags: there are only 413 PROFESSION, 206 LOCATION-ORGANIZATION and 17 LOCATION-OTHER tags in the entire i2b2/UTHealth shared task corpus [5], and the tags do not exist in other de-identified corpora, such as the i2b2 2006 challenge data [16].

7.1. Comparison to 2006 i2b2 de-identification shared task

The existence of the 2006 i2b2 de-identification challenge [16] raises the question of whether de-identification systems have improved significantly in the past eight years. However, differences in the data sets, annotation schemes, and evaluation software make the comparison between participating systems somewhat tricky.

To begin with, the data from the 2006 shared task were tokenized before the organizers shared it with the task participants. However, for the 2014 data we chose to not make any such modifications, preferring instead to share the data in the same form they were found in the Partners data repository. The lack of tokenization makes the 2014 task more difficult, and somewhat changes the evaluation metric, as the "token-based" evaluation simply uses whitespace to determine tokenization, rather than using an automated tokenizing system. The 2006 data consisted of 889 discharge summaries (669 training, 220 test), while the

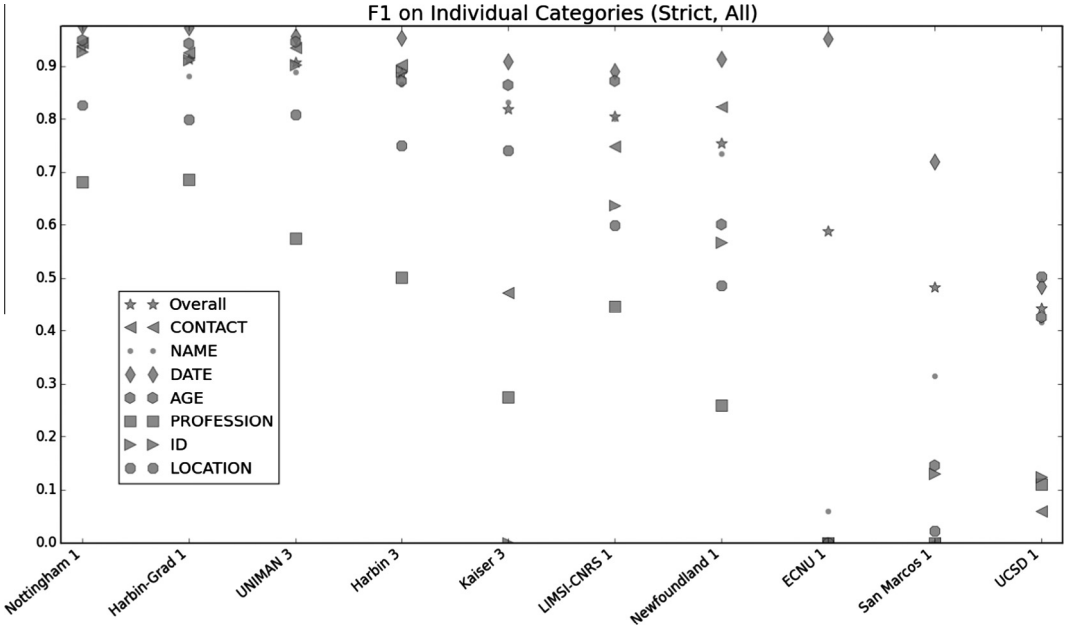

Fig. 1. Micro-averaged entity-based micro-averaged F_1 measures by category: i2b2-PHI categories.

Table 7
Comparison of 2006 and 2014 team results: precision, recall, and F_1 of token-based binary HIPAA-PHI. Top result for each team selected by F_1.

2006 teams				2014 teams			
Team name/ run	P	R	F_1	Team name/ run	P	R	F_1
Wellner 3	98.7	97.5	.981	Nottingham 1	.9900	.9640	.9768
Szarvas 2	99.3	96.9	.980	Harbin-Grad 2	.9776	.9629	.9702
Aramaki 1	99.1	94.9	.970	Manchester 3	.9825	.9567	.9694
Hara 3	96.1	93.8	.949	Harbin 3	.9747	.9408	.9575
Wrenn 3	94.9	94.3	.946	Kaiser 3	.9692	.8999	.9333
Guo 1	93.8	88.2	.909	Newfoundland 1	.9207	.8522	.8851
Guillen 1	92.8	70.5	.801	LIMSI–CNRS 3	.9605	.8048	.8758
				ECNU 1	.9666	.7404	.8385
				San Marcos 1	.9095	.6045	.7263
				UCSD 1	.7591	.5161	.6144

2014 data contained a wider variety of clinical records, including discharge summaries, admission notes, and correspondences between doctors.

Additionally, the 2006 annotation scheme was based more closely on the HIPAA-PHI categories, using only PATIENT, DOCTOR, LOCATION, HOSPITAL, DATE, ID, and PHONE. In part, the smaller number of categories in the 2006 data was due to the lack of any examples of the other PHI categories, such as fax numbers and emails. As shown in the earlier sections, the 2014 data contained some PHI categories with less representation (such as PROFESSION and ORGANIZATION). The 2006 data annotated only the day and month of dates, while the 2014 data annotated all parts of dates, including years.

Finally, during surrogate generation, the 2006 data included ambiguous terms (substituting procedure and device names for people and locations) and out-of-vocabulary terms (i.e., deliberately introduced misspellings). While the 2014 data do include misspellings, they do not introduce deliberately ambiguous terms as PHI.

Despite these differences, the two tasks are similar enough that we can still perform some basic comparisons. We performed binary (PHI vs non-PHI) entity- and token-based evaluations on the 2014 data, looking only at the HIPAA-PHI categories, for this purpose.

Table 8
Comparison of 2006 and 2014 team results: precision, recall, and F_1 of entity-based binary HIPAA-PHI.

2006 teams				2014 teams			
Team name/ run	P	R	F_1	Team name/ run	P	R	F_1
Szarvas 3	.978	.957	.967	Nottingham 1	.9776	.9403	.9586
Wellner 3	.967	.959	.963	Harbin-Grad 1	.9557	.9350	.9452
Aramaki 1	.951	.931	.941	Manchester 3	.9459	.9234	.9345
Hara 3	.910	.896	.903	Harbin 3	.9459	.9000	.9224
Wrenn 3	.887	.918	.902	Kaiser 3	.8919	.8076	.8477
Guo 1	.87	.798	.813	LIMSI–CNRS 3	.9219	.7736	.8413
				Newfoundland 1	.8628	.7654	.8112
				ECNU 1	.9342	.6122	.7397
				San Marcos 1	.7780	.4584	.5769
				UCSD 1	.5403	.3914	.4540

Table A1
Participants in Track 1 of the 2014 i2b2/UTHealth NLP shared task.

Team name	Affiliations	# of members	Countries
Nottingham	University of Nottingham	2	UK
UNIMAN	University of Manchester University of Novi Sad Health eResearch Centre	5	UK Serbia
Harbin	Harbin Institute of Technology	5	China
Harbin-Grad	Harbin Institute of Technology Shenzhen Graduate School	5	China
ECNU	East China Normal University	1	China
Kaiser	Kaiser Permanente Southern California	7	USA
LIMSI–CNRS	Centre National de la Recherche Scientifique	1	France
San Marcos	California State University San Marcos	1	USA
Newfoundland	Memorial University of Newfoundland	1	Canada
UCSD	University of California San Diego	5	USA

Table A2
A set of variations of system output, and how they would be compared to the gold standard using the different levels of evaluation. A + indicates that the evaluation method would result in a true positive, a – indicates that the input would result in a false positive.

Gold standard:<LOCATION type="HOSPITAL">Brooks Infirmary</LOCATION>	Instance-based			Token-based		
	i2b PHI categories and types	HIPAA PHI categories only	PHI/not PHI	i2b2 PHI categories and types	HIPAA PHI categories only	PHI/not PHI
<LOCATION type="HOSPITAL">Brooks Infirmary</LOCATION>	+	+	+	+	+	+
<LOCATION type="ORGANIZATION">Brooks Infirmary</LOCATION>	–	+	+	+	+	+
<NAME type="DOCTOR">Brooks Infirmary</NAME>	–	–	+	–	–	+
<LOCATION type="HOSPITAL">Brooks</LOCATION>	–	–	–	+	+	+
<LOCATION type="ORGANIZATION">Brooks</LOCATION>	–	–	–	–	+	+
<NAME type="DOCTOR">Brooks</NAME>	–	–	–	–	–	+

Table 7 shows the best micro-averaged token-based P, R, and F_1 scores from the top run for each 2006 and 2014 team using the binary evaluation on only the HIPAA-PHI categories. Best team run was selected based on F_1. Table 8 shows the same information at the entity level.

For P, R, and F_1, both the token-based and entity-based comparisons show that the top 2006 systems perform slightly better, though the differences are relatively small. Given the aforementioned differences between the two corpora, we can conclude that overall the systems from 2014 are at least on par with the 2006 systems.

8. Discussion

In the overview paper for the 2006 i2b2 de-identification shared task, the authors posed the following questions: "1. Does success on this challenge problem extrapolate to similar performance on other, untested data sets? 2. Can health policy makers rely on this level of performance to permit automated or semi-automated de-identification of health data for research purposes without undue risk to patients?" [16].

In general, it remains difficult to say whether the systems built for this challenge will perform as well on other data. While the data for the 2014 shared task included a wider variety of document types than the 2006 data, both sets were drawn from the Partners HealthCare and so share a certain degree of similarity. In order to truly determine if the performance will extrapolate to other data sets, we will need data sets from other medical institutions that have PHI identified and replaced with surrogates in a similar fashion.

The answer to the second question is similarly difficult to determine, for similar reasons. However, for institutions that use a data format similar to that of Partners, the answer could be positive. While we are not aware of an industry-wide standard, 95% has been suggested as a rule-of thumb for determining whether a system can reliably de-identify a data set for safe distribution. Table 2 shows that, looking at HIPAA-PHI only and using a token-based evaluation, the top 4 systems satisfy this requirement. A related consideration is whether perfect de-identification (100% precision and recall) is a realistic goal. Given the performance of the participating systems in this challenge, as well as other recently developed de-identification software, it may be that perfect de-identification is unachievable, with the best performances we can expect being around .95 or slightly higher.

9. Conclusion

This paper presents an overview of the de-identification track (Track 1) from the 2014 i2b2/UTHealth NLP shared task. Due to the different needs differing institutions might have for de-identifying records, this task investigates performance on de-identification at both entity and token levels, for various definitions of PHI: i2b2-PHI, which match the gold standard; HIPAA-PHI, which adhere strictly to the HIPAA guidelines for de-identification; and binary PHI, which consider only whether a PHI is identified as PHI at all. Of these, the entity-based i2b2-PHI de-identification was the most difficult, with the highest-ranked team achieving a micro-averaged F_1 of 0.9360.

In its most strict form, de-identification remains a task that cannot yet be handled perfectly by automated systems; however, the performances of the systems are encouraging and can solve a significant portion of the task. Whether this performance is "good enough" remains a topic of debate and depends on the PHI types that are missed (e.g., doctor names vs patient names would have different significance for perfect identification). Until these debates are resolved, we expect most data will be distributed with data use agreements that tackle the problem from the policy end, thus strengthening the solutions provided by the automated systems.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We would like to thank the program committee for the 2014 i2b2/UTHealth NLP Shared Task, along with everyone who participated in the task and workshop. We would also like to that the JBI editor and reviewers, for their thoughtful comments and feedback.

Funding for this project was provided by:

- NIH NLM 2U54LM008748, PI: Isaac Kohane.
- NIH NLM 5R13LM011411, PI: Özlem Uzuner.
- NIH NIGMS 5R01GM102282, PI: Hua Xu.

Appendix A

See Tables A1 and A2.

References

[1] A. Stubbs, C. Kotfila, H. Xu, Ö. Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inf. 58S (2015) S67–S77.
[2] K. El Emam, D. Buckeridge, R. Tamblyn, A. Neisa, E. Jonker, A. Verma, The re-identification risk of Canadians from longitudinal demographics, BMC Med. Inf. Decision Making 11 (2011) 46.
[3] P. Golle, Revisiting the uniqueness of simple demographics in the US population, Workshop on Privacy in the Electronic Society, 2006.

[4] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4, Pittsburgh, 2000.

[5] Amber Stubbs, Özlem Uzuner, Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus, J. Biomed. Inf. 58S (2015) S20–S29.

[6] A. Stubbs, Ö. Uzuner, C.Kotfila, I. Goldstein, P. Szolovitz, Challenges in synthesizing replacements for PHI in narrative EMRs, in: Aris Gkoulalas-Divanis, Grigorios Loukides (Eds.), Medical Data Privacy Handbook, Springer, 2015.

[7] B. Sundheim, Overview of results of the MUC-6 evaluation, in: Proceedings of the Sixth Message Understanding Conference (MUC-6), 1995.

[8] Nancy Chinchor, MUC-7 Named Entity Task Definition version 3.5, 1997. <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html> (last accessed 06.01.15).

[9] Nancy Chinchor (a), Named Entity Scores – English, in: Message Understanding Conference Proceedings, Updated January 12, 2001. <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_english_score_report.html> (last accessed 06.01.15).

[10] L. Hirschman, A. Yeh, C. Blaschke, A. Valencia, Overview of BioCreAtIvE: critical assessment of information extraction for biology, BMC Bioinf. 6 (2005) s1.

[11] William Hersh, Ellen Voorhees, TREC genomics special issue overview, Inform. Retrieval 12 (1) (2009) 1–15.

[12] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, Jun'ichi Tsujii, Overview of BioNLP'09 shared task on event extraction, in: Proceedings of the Workshop on BioNLP: Shared Task, Association for Computational Linguistics, Boulder, Colorado, June 2009, pp. 1–9.

[13] S. Pyysalo, T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, S. Ananiadou, Overview of the ID, EPI and REL tasks of BioNLP shared task 2011, BMC Bioinf. 13 (2012).

[14] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, P. Zweigenbaum, in: Overview of BioNLP shared task 2013 Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, 2013, pp. 1–7.

[15] BioASQ project, Data. <http://www.bioasq.org/participate/data> (last accessed 07.01.15).

[16] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.

[17] Stephane M. Meystre, Friedlin F. Jeffrey, Brett R. South, Shuying Shen, Matthew H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (1) (2010) 70, http://dx.doi.org/10.1186/1471-2288-10-70.

[18] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, Lynette Hirschman, The MITRE identification scrubber toolkit: design, training, and assessment, Int. J. Med. Informatics 79 (12) (2010) 849–859, http://dx.doi.org/10.1016/j.ijmedinf.2010.09.007.

[19] Oscar Ferrández, Brett R. South, Shuying Shen, F. Jeffrey Friedlin, Matthew H. Samore, Stéphane M. Meystre, BoB, a best-of-breed automated text de-identification system for VHA clinical documents, J. Am. Med. Inf. Assoc. 20 (1) (2013) 77–83, http://dx.doi.org/10.1136/amiajnl-2012-001020.

[20] Louise Deleger, Todd Lingren, Yizhao Ni, Meghan Kaiser, Laura Stoutenborough, Keith Marsolo, Michal Kouril, Katalin Molnar, Imre Solti, Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research, J. Biomed. Inf. 50 (2014) 173–183, http://dx.doi.org/10.1016/j.jbi.2014.01.014.

[21] Benjamin Wellner, Sequence Models and Ranking Methods for Discourse Parsing, Ph.D. Dissertation, Brandeis University, Waltham, MA, 2009.

[22] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. 18th International Conference on Machine Learning, 2001.

[23] Apache UIMA, 2006. <http://uima.apache.org> (accessed May 2015).

[24] G.K. Savova, J.J. Masanz, P.V. Ogren, et al., Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.

[25] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92, 1992, p. 144. http://dx.doi.org/10.1145/130385.130401.

[26] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, et al., Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, J. Am. Med. Inform. Assoc. (2012), http://dx.doi.org/10.1136/amiajnl-2012-001012.

[27] A.C. McCallum, MALLET: a Machine Learning for Language Toolkit; 2002.

[28] Vishesh Kumar, Amber Stubbs, Stanley Shaw, Özlem Uzuner, Creation of a new longitudinal corpus of clinical narratives, J. Biomed. Inf. 58S (2015) S6–S10.

[29] A. Stubbs, Ö. Uzuner, De-identification of medical records through annotation, in: Nancy Ide, James Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer, Anticipated Publication, 2015.

[30] Nancy Chinchor, The statistical significance of the MUC-4 results, in: Proceedings of the 4th Conference on Message Understanding, 1992, pp. 30–50.

[31] E.W. Noreen, Computer-Intensive Methods for Testing Hypotheses: An Introduction, Wiley, New York, 1989.

[32] Bin He, Yi Guan, Jianyi Cheng, Keting Cen, Wenlan Hua, CRFs based de-identification of medical records, J. Biomed. Inf., 2014 (i2b2 NLP supplement) J. Biomed. Inform. 58S (2015) S39–S46.

[33] Zengjian Liu, Yangxin Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Haodi Li, Jingfeng Wang, Qiwen Deng, Suisong Zhu, Automatic de-identification of electronic medical records using token-level and character-level conditional random fields, J. Biomed. Inf.: 2014 (i2b2 NLP supplement) J. Biomed. Inform. 58S (2015) S47–S52.

[34] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inf. Assoc.: JAMIA 17 (1) (2010) 19–24, http://dx.doi.org/10.1197/jamia.M3378.

[35] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T. Wiley, Daniel Zisook, Yang Huang, De-identification and risk factor detection in medical records, Paper presented at the Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data. November 14, 2014, Washington DC. J. Biomed. Inform. 58S (2015) S164–S170.

[36] Cyril Grouin, Clinical records de-identification using CRF and rule-based approaches, Poster session, presented at the Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data. November 14, 2014, Washington DC.

[37] Azad Dehghan, Aleksandar Kovačević, George Karystianis, John A. Keane, Goran Nenadic, Combining knowledge- and data-driven methods for de-identification of clinical narratives, J. Biomed. Inf., 2014 (i2b2 NLP supplement) J. Biomed. Inform. 58S (2015) S53–S59.

[38] Tao Chen, Hidden Markov model using Dirchilet process for de-identification, J. Biomed. Inf., 2014 (i2b2 NLP supplement) J. Biomed. Inform. 58S (2015) S60–S66.

[39] P. Orbanz, Y.W. Teh, Bayesian nonparametric models, Encyclopedia of Machine Learning, Springer, 2010.

[40] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (6) (1966) 1554–1563.

[41] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent dirichlet allocation, J. Mach. Learning Res. 3 (2003) 993–1022.

[42] Hui Yang, Jonathan Garibaldi, Automatic detection of protected health information from clinic narratives, J. Biomed. Inf., 2014 (i2b2 NLP supplement) J. Biomed. Inform. 58S (2015) S30–S38.

[43] Rocio Guillen, An approach to de-identifying electronic medical records, Poster session, presented at the Seventh i2b2 Shared Task and Workshop: Challenges in Natural Language Processing for Clinical Data. November 14, 2014, Washington DC.