Task 1 of the CLEF eHealth Evaluation Lab 2016: Handover Information Extraction

Hanna Suominen[1], Liyuan Zhou[2], Lorraine Goeuriot[3], and Liadh Kelly[4]*

[1] Data61, The Australian National University (ANU), University of Canberra, and University of Turku, Canberra, ACT, Australia,
hanna.suominen@data61.csiro.au

[2] Data61 and ANU, Canberra, ACT, Australia,
liyuan.zhou@data61.csiro.au

[3] LIG, Université Grenoble Alpes, France, Lorraine.Goeuriot@imag.fr

[4] ADAPT Centre, Trinity College, Dublin, Ireland Liadh.Kelly@tcd.ie

Abstract. Cascaded speech recognition (SR) and information extraction (IE) could support the best practice for clinical handover and release clinicians' time from writing documents to patient interaction and education. However, high requirements for processing correctness evoke methodological challenges and hence, processing correctness needs to be carefully evaluated as meeting the requirements. This overview paper reports on how these issues were addressed in a shared task of the eHealth evaluation lab of the Conference and Labs of the Evaluation Forum (CLEF) in 2016. This IE task built on the 2015 CLEF eHealth Task on SR by using its 201 synthetic handover documents for training and validation (appr. $8,500 + 7,700$ words) and releasing another 100 documents with over $6,500$ expert-annotated words for testing. It attracted 25 team registrations and 3 team submissions with 2 methods each. When using the macro-averaged *F1* over the 35 form headings present in the training documents for evaluation on the test documents, all participant methods outperformed all 4 baselines, including the organizers' method ($F1 = 0.25$), published in 2015 in a top-tier medical informatics journal and provided to the participants as an option to build on, a random classifier ($F1 = 0.02$), and majority classifiers for the two most common classes (i.e., NA to filter out text irrelevant to the form and the most common form heading, both with $F1 < 0.00$). The top-2 methods ($F1 = 0.38$ and 0.37) had statistically significantly ($p < 0.05$, Wilcoxon signed-rank test) better performance than the third-best method ($F1 = 0.35$). In comparison, the top-3 methods and the organizers' method (7th) had *F1* of 0.81, 0.80, 0.81, and 0.75 in the NA class, respectively.

Keywords: Computer Systems Evaluation, Data Collection, Information Extraction, Medical Informatics, Nursing Records, Patient Handoff, Patient Handover, Records as Topic, Software Design, Speech Recognition, Test-set Generation, Text Classification

* HS designed and led the task as a part of the CLEFeHealth2016 evaluation lab, chaired by LG and LK. Together with LZ, she developed its data sets and evaluation methodology. HS drafted the paper and after this all authors revised and approved it.

1 Introduction

Fluent information flow, defined as channels, communication, contact, or linkages to pertinent people [1], is critical in healthcare in general and in particular in *clinical handover* or *handoff*, where a nurse, other clinician, or a group of clinicians is transferring professional responsibility and accountability, for example, when changing shifts [2]. This *shift-change nursing handover* is a form of clinical narrative where only a small part of the flow is documented *electronically in writing* [3] although without this, anything from two-thirds to all *spoken* handover information transferred incorrectly or lost completely already after a couple of shift changes [4, 5]. Moreover, failures in the flow of information from nursing handover are a major contributing factor in over two-thirds of sentinel events in hospitals and associated with over a tenth of preventable adverse events [2].

As a mechanism to contribute to quality and safety in healthcare, best practice for the handover information documentation recommends standardized, structured, and synchronous processes at patient site in the presence *and* active involvement of the patients, and where relevant, their next-of-kin [6].[5] In order to support their compliance, cascaded *speech recognition* (SR) with *information extraction* (IE) has been studied since 2015 [7, 8], also as part of the *eHealth evaluation lab* by the *Conference and Labs of the Evaluation Forum* (CLEF) [9]. As justified empirically in clinical settings in 2014, the cascade pre-fills a structured handover form for a clinician to proof, edit, and sign off [10, 11]. Based on the rate of information loss above, the approach of the nurse who is handing over revising and signing off the document draft him/herself any time before the shift ends (but preferably immediately after the handover) can decrease the loss from 0 to 13 per cent.

This novel application evokes fruitful challenges for method research and development, and consequently, its SR part and IE parts were chosen as the *CLEF eHealth 2015 Task 1a* [12] and *CLEF eHealth 2016 Task 1* [13], respectively.[6] First, SR is complicated by clinical characteristics of a large number of nurses moving between patient sites to involve patients in handover, resulting in a noisy minimally-personalized multi-speaker setting far from a typical case with a single person, equipped with a personalized SR engine, speaking in a peaceful office. Second, SR errors multiply in cascading with IE. Because of the severe implications that these errors may have in clinical judgment and decision-making, the cascade correctness needs to be carefully evaluated as meeting the healthcare requirements.

The cascaded tasks align with the CLEF eHealth usage scenario of easing patients, their next-of-kin, and other *laypersons* in understanding and accessing *electronic health* (eHealth) information [14, 15]. Namely, the application could

[5] See also similar guidance by the *World Health Organisation* (WHO) at `http://www.who.int/patientsafety/research/methods_measures/human_factors/organizational_tools/en/` (last accessed on 1 July 2016).

[6] See `https://sites.google.com/site/clefehealth2015/` and `https://sites.google.com/site/clefehealth2016/` (last accessed on 1 July 2016).

release a substantial amount of healthcare workers' time from documentation to, for example, longer discussions about the findings, care plans, and consumer-friendly resources for further information with the patients, and/or their next-of-kin. Fulfilling the legal requirement to document every event in healthcare can take over half of nurses' working time with centralized clinical information systems or fully structured information entry (whilst free-form text entry at the patient-site decreases this to a few minutes per patient) [16–18]. SR drafts a written document from a tenth to three-quarters of the time it takes to transcribe this by hand, whilst the clinician's proofing time is approximately the same in both cases [19], or equivalently, draft text for a minute of handover speech (with 160 words, corresponding to the range that people comfortably hear and vocalize words [20]) is available only 20 seconds after finishing the handover with a real-time SR engine that recognizes at least as many words per minute as a very skilled typist (i.e., 120 [21]). Cascading this with content structuring through IE can bring further efficiency gains by easing finding information and making this content available for computerized surveillance and decision-making in healthcare [22].

The rest of this organizers' overview of the Task 1 of the CLEF eHealth Evaluation Lab 2016 on handover IE is organized as follows: In Section 2, we describe the materials and methods provided and used by the organizers in the task. In Section 3, we introduce the task results. In Section 4, we conclude by summarizing the main findings, relating them with previous work on clinical IE, and discussing the significance of this study.

2 Materials and Methods

2.1 Text Documents

The *NICTA Synthetic Nursing Handover Data* [8, 9] was used in Task 1.[7] This set of 101 synthetic patient cases for training, another 100 validation, and yet another 100 for testing was developed for SR and IE related to nursing shift-change handover in 2012–2016. The dataset was authored by a *registered nurse* (RN) with over 12 years' experience in clinical nursing and its content was thus very similar to real documents in Australian English (which cannot be made available). Each case consisted of a *patient profile*, a written, free-form text paragraph (i.e., the *written handover document*); and for SR purposes, its spoken (i.e., the *verbal handover document*) and*speech-recognized counterparts*.

The written handover documents were used in Task 1 with the training, validation, and test set having 8, 487, 7, 730, and 6, 540 words in total. In the last year's Task 1a on SR, the 101 training and 100 validation cases were used for training and testing; for this year's Task 1 on IE, the dataset was supplemented with another independent test set of 100 cases.

[7] See `https://www.nicta.com.au/nicta-synthetic-nursing-handover-open-data-software-and-demonstrations/` (last accessed on 1 July 2016).

The data releases with the requirement to cite [8] for the training set, [9] for the validation set, and this paper for the test set were approved at NICTA and the RN was consented in writing. The spoken, free-form text documents were licensed under *Creative Commons — Attribution Alone — Non-commercial — No Derivative Works* (CC-BY-NC-ND) and the remaining documents under *Creative Commons — Attribution Alone* (CC-BY).

2.2 Human Annotations

In Task 1, the written handover documents were annotated, by the aforementioned RN using the *Protégé 3.1.1 Knowtator 1.9 beta* [23], with respect to a form with 49 headings (aka classes) to fill out. The form was compatible with existing handover forms, matched the Australian and international standards and best practice for handover communication, and mimicked the RN's practical experiences from two Australian states/territories [7, 8].[8]

Alphabetically, the following 35 of these classes were present in the training set:

- Appointment/Procedure: 1) City, 2) ClinicianGivenNames/Initials, 3) ClinicianLastname, 4) Day, 5) Description, 6) Status, 7) Time, and 8) Ward,
- Future: 9) Alert/Warning/AbnormalResult, 10) Discharge/TransferPlan, and 11) Goal/TaskToBeCompleted/ExpectedOutcome,
- Medication: 12) Dosage, 13) Medicine, and 14) Status,
- MyShift: 15) ActivitiesOfDailyLiving, 16) Contraption, 17) Input/Diet, 18) OtherObservation, 19) Output/Diuresis/BowelMovement, 20) RiskManagement, 21) Status, and 22) Wounds/Skin, and
- PatientIntroduction: 23) AdmissionReason/Diagnosis, 24) Ageinyears, 25) Allergy, 26) CarePlan, 27) ChronicCondition, 28) CurrentBed, 29) CurrentRoom, 30) Disease/ProblemHistory, 31) Gender, 32) GivenNames/Initials, 33) Lastname, 34) UnderDr_GivenNames/Initials, and 35) UnderDr_Lastname.

Irrelevant text was to be classified as NA and the annotation task was seen as *multi-class classification*, that is, each word could belong to precisely one class.

To improve the annotation consistency in including/excluding articles or titles and in marking gender information in each document if it was available, some light proofing was performed semi-automatically by HS and LZ before releasing the classification *gold standard* (GS) under the CC-BY license.

2.3 Measures in Performance Evaluation

Precision (Prec), *Recall* (Rec), and their harmonic mean

$$\mathit{F1} = \frac{2\ \text{Prec Rec}}{\text{Prec} + \text{Rec}}$$

[8] For further information on and illustration of the dataset, its creation, and the form, we refer the reader to the Methods section of our previous paper [8].

were measured. Performance was evaluated first separately in every heading and NA. That is, if TP_c, FP_c, and FN_c refer to the *numbers of true positives*, *false positives*, and *false negatives* for a class $c \in \{1, 2, 3, \ldots, 35, 36 = \text{NA}\}$, respectively, the class-specific measures were defined as

$$\text{Prec}_c = \frac{TP_c}{TP_c + FP_c}, \text{Rec}_c = \frac{TP_c}{TP_c + FN_c}, \text{ and}$$

$F1_c$ as their harmonic mean. Then, we documented the performance in the dominant class of 36 = NA and averaged over the first 35 classes present in the training set by using *macro-averaging* (MaA) and *micro-averaging* (MiA) with the former measures being

$$\text{Prec}_{\text{MaA}} = \frac{\sum_{c=1}^{35} \text{Prec}_c}{35}, \text{Rec}_{\text{MaA}} = \frac{\sum_{c=1}^{35} \text{Rec}_c}{35}, \text{ and}$$

$F1_{\text{MaA}}$ their harmonic mean and the latter measures being

$$\text{Prec}_{\text{MiA}} = \frac{\sum_{c=1}^{35} TP_c}{\sum_{c=1}^{35} TP_c + FP_c}, \text{Rec}_{\text{MiA}} = \frac{\sum_{c=1}^{35} TP_c}{\sum_{c=1}^{35} TP_c + FN_c}, \text{ and}$$

$F1_{\text{MiA}}$ their harmonic mean.

Because our desire was to perform well in all classes, and not only in the majority classes, the macro-averaged results were to be emphasized over the micro-averaged results. Hence, this $F1_{\text{MaA}}$ was used to rank the participant submissions. The 14 validation words and 27 test words annotated with a class not present in the training set were excluded from the evaluation.

2.4 Baselines Methods in Performance Evaluation

This year we were aiming to lower the entry barrier and encourage novelty in Task 1 by providing participants with not only an evaluation script (i.e., the *CoNLL 2000 Shared Task on Chunking*[9]) but also processing code for IE, together with all its intermediate and final outputs from our previous paper [8]. This organizers' method called *NICTA* served as one of the four *baseline methods*.

The NICTA method solved the IE task by using the *CRF++* implementation[10] of the *Conditional Random Fields* [24] trained on the training set and validated on the validation set [25, 26], prior to testing it on the independent test set. The method generated its eight *syntactic* (e.g., the lemma, part of speech tag, and parse tree of a given word), three *semantic* (i.e., the top-5 candidates of a given word retrieved the *Unified Medical Language System* (UMLS), its top UMLS mapping, and its medication score, derived from the *Anatomical Therapeutic Chemical List*), and two *statistical feature types* (i.e., the location of a

[9] See `http://www.cnts.ua.ac.be/conll2000/chunking/` (last accessed on 1 July 2016).

[10] See `http://taku910.github.io/crfpp/` (last accessed on 1 July 2016).

given word on a ten-point scale from the beginning of the document to its end and the number of times a given term occurs in a document divided by the maximum of this term frequency over all terms in the document) by processing the original text documents using *Stanford CoreNLP* (English grammar) by the *Stanford Natural Language Processing Group* [27], *MetaMap 2012* by the *US National Library of Medicine* [28], and *Ontoserver* by the Australian *Commonwealth Scientific and Industrial Research Organisation* (CSIRO) [29].

The other three baselines were *Random* (i.e., classifying each word by selecting one out of the 36 classes randomly[11]), *NA* [i.e., classifying each word as belonging to the dominant training class of 36) NA], and *Majority* [i.e., classifying each word as belonging to the majority training class of 11) Future_Goal/TaskToBeCompleted/ExpectedOutcome].

2.5 Statistical Significance Testing in Performance Evaluation

Statistical differences between the $F1_{\mathrm{MaA}}$ percentages of the methods were evaluated in Task 1 using the *R 3.2.4* implementation of the *Wilcoxon signed-rank test* (W) [30].[12] This test was chosen as an alternative to the paired t-test, because of not being able assume that the $F1_{\mathrm{MaA}}$ percentages for the sample of the 100 test documents were normally distributed.

After ranking the baselines and submissions based on their $F1_{\mathrm{MaA}}$ on the entire test set, W was computed for the paired comparisons from the best and second-best method to the second-worst and worst method. The resulting p value and the significance level of 0.05 was used to determine if the median performance of the higher-ranked method was significantly better than this value for the lower-ranked method.

3 Results

The task released a training set of 101 synthetic clinical documents on 30 October 2015; an independent validation set of 100 documents on 30 October 2015; and an independent test set of 100 documents on 15 April 2016. The test set annotations were not released before 1 August 2016.

The task was open for everybody. We particularly welcomed academic and industrial researchers, scientists, engineers and graduate students in SR, natural language processing, and biomedical/health informatics. We also encouraged participation by multi-disciplinary teams that combine technological skills with content expertise in nursing.

By 30 April 2016, 25 teams had registered their interest in the task through the *CLEF 2016 registration system*.[13] Each team was allowed to submit two fully automated methods (or compilations) — referred to as using the suffixes$A$ and B) after the team name.

[11] using `https://www.random.org/`, last accessed on 1 July 2016

[12] See `http://www.r-project.org/` (last accessed on 1 July 2016).

[13] See `http://clef2016.clef-initiative.eu` (last accessed on 1 July 2016).

By 1 May 2016, regardless of the difficulty of 36-class classification with only about $16,200$ training and validation instances in total, three teams had submitted two IE methods each. The team *TUC-MI* (that also participated the 2015 SR task) originated from Germany, *LQRZ* from the Netherlands, and *ECNU_ICA* from China.

The team TUC-MI followed an interdisciplinary approach and consisted of four computer scientists, supervised by one professor. Two scientists brought the expertise from the field of natural language processing as well as information retrieval exploring features for the clinical context. The other two scientists had practical experience in machine learning and computational linguistics. The latter group developed strategies for feature subset selection and performed parameter optimization for learning methods. The team's approach focused on the exploration of relevant features for CRFs. Therefore, the team used wrappers for feature subset selection in conjunction with parameter optimization to consider how the algorithm and the dataset interact. First, as TUC-MI-A, the team composed a set of 41 features based on Stanford CoreNLP, latent Dirichlet allocation, regular expressions, and the ontologies of WordNet and UMLS. Next, the team applied the heuristic methods best-first and greedy (hill-climbing) with forward and backward direction to feature evaluation and selection. In the development phase, the team also observed that 19 out of 41 features performed best in combination with the hyperparameter $C = 10$ of the CRF classifier and submitted this configuration as TUC-MI-B.

The team LQRZ had an Artificial Intelligence Master Student, a Postdoc supervisor, and a Master in Artificial Intelligence supervisor. Its both methods tried to avoid the feature engineering handcraft effort and aimed at using general domain data (no clinical specifics). The method LQRZ-A consisted of a one-hidden-layer (tanh) Multilayer Perceptron, with a context window of 7, trained with Adagrad for 50 epochs. Its parameters were the following: $W1$: word embeddings — intersected with the GoogleNews pretrained embeddings (300 dimensions), $b1$: uniform bias, $W2$: uniform weights, and $b2$: uniform bias. The output layer (softmax) computed the probability of the word belonging to each form tag (39 tags, considering the training and validation set tags). The method LQRZ-B consisted of an ensemble method. On the first step, a Random forest was used to identify NA tags (by binary-discriminating between NA and Others). On the second step, a one-hidden-layer (tanh) Multilayer Perceptron, with a context window of 7 was used to categorize between the remaining tags. The Random forest used the GoogleNews pretrained embeddings (300 dimensions) as features (and a random-sampled vector if the word was not found in the word2vec model). The Multilayer Perceptron parameters were as above.

The team ECNU_ICA had six people in total: a PhD student, three graduate students, and two professors. They had participated evaluation labs on clinical language processing before. In its method ECNU_ICA-A, the team extracted the bed number, room number, age, and doctor's name by using rules. Then, the team ran the organizers' CRF and combined its result with the result obtained by utilizing rules. In contrast, in the method ECNU_ICA-B, the team selected

Table 1. Performance of the four baselines and six submissions on the 100 test documents. In the $F1_{\mathrm{MaA}}$ column, an asterisk (*) indicates that the method of the row was significantly better than the method of the row below ($p < 0.05$, Wilcoxon signed-rank test). The top-3 values for each measure are *emphasized*. When the organizers' announced the evaluation results on the test set to the participants, the team LQRZ noticed an unfortunate human error in the output writing of both methods. The organizers allowed them to re-submit and the results corresponding to the corrected outputs are included below.

Method	Macro-averaged			Micro-averaged			Class NA		
	Prec	Rec	*F1*	Prec	Rec	*F1*	Prec	Rec	*F1*
TUC-MI-B	_0.493_	_0.369_	_0.382_	0.500	_0.505_	0.503	_0.812_	0.802	_0.807_
ECNU_ICA-A	_0.493_	_0.406_	_0.374_*	_0.510_	_0.522_	_0.516_	_0.816_	0.788	_0.802_
LQRZ-B	0.425	_0.383_	_0.345_	0.490	_0.517_	0.503	_0.849_	0.779	_0.813_
TUC-MI-A	0.423	0.300	0.311	0.503	0.443	0.471	0.726	0.850	0.783
LQRZ-A	0.411	0.307	0.308	_0.563_	0.472	_0.514_	0.723	_0.894_	0.800
ECNU_ICA-B	0.428	0.292	0.297*	_0.581_	0.459	_0.513_	0.675	_0.881_	0.764
NICTA	_0.435_	0.233	0.246*	0.433	0.368	0.398	0.682	0.831	0.749
Random	0.018	0.028	0.019*	0.018	0.030	0.022	0.405	0.030	0.055
Majority	0.000	0.029	0.001	0.016	0.027	0.020	0.000	0.000	0.000
NA	0.000	0.000	0.000	0.000	0.000	0.000	0.407	_1.000_	0.579

the best suitable feature set for each class from the feature types provided by the organizers. Then, the team ran the organizers' CRF based on those features only. At last, it use the method of voting to determine the class for each word, used the IE rules of the ECNU_ICA-A for the bed number, room number, age, and doctor's name, and combined the results obtained by utilizing rules and voting methods.

We organizers' were excited that all three participating teams scored among the top 3 and all six participant methods outperformed all four baselines in Task 1 (Tables 1 and 2), including the organizers' method, published in 2015 in a top-tier medical informatics journal. The top-2 methods had the $F1_{\mathrm{MaA}}$ percentages of 38.2 (TUC-MI-B with $F1_{\mathrm{NA}}$ of 80.7 per cent) and 37.4 (ECNU_ICA-A with $F1_{\mathrm{NA}}$ of 80.2 per cent), respectively. Their difference was not statistically significant but they were significantly better than the 34.5 per cent performance of the third-best method (LQRZ-B with $F1_{\mathrm{NA}}$ of 81.3 per cent). In comparison, the NICTA baseline with its $F1_{\mathrm{MaA}}$ percentage of 24.6 (and $F1_{\mathrm{NA}}$ of 74.9 per cent) was significantly worse than the participant methods but significantly better than the Random, Majority, and NA baselines with the respective $F1_{\mathrm{MaA}}$ percentages of 1.9, 0.1, and 0.0.

4 Discussion

The macro-averaged *F1* scores of the top-3 methods over the 35 form headings (i.e., 0.382, 0.374, and 0.345) demostrate the great difficulty in performing well for each heading. However, they all and even the organizers' NICTA baseline

Table 2. Performance on the 101 training and 100 validation documents

Method	Macro-averaged			Micro-averaged			Class NA		
	Prec	Rec	*F1*	Prec	Rec	*F1*	Prec	Rec	*F1*
Training									
TUC-MI-B	0.998	0.995	0.996	0.997	0.998	0.997	0.998	0.997	0.998
ECNU_ICA-A	0.995	0.992	0.994	0.995	0.991	0.993	0.993	0.998	0.995
LQRZ-B	0.768	0.699	0.718	0.810	0.861	0.835	0.859	0.791	0.824
TUC-MI-A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LQRZ-A	0.741	0.591	0.624	0.908	0.862	0.884	0.920	0.979	0.948
ECNU_ICA-B	0.454	0.328	0.344	0.461	0.528	0.492	0.864	0.706	0.777
NICTA	1.000	0.976	0.980	1.000	0.914	0.955	0.903	1.000	0.949
Random	0.017	0.027	0.017	0.017	0.030	0.022	0.490	0.032	0.060
Majority	0.002	0.029	0.003	0.058	0.105	0.075	0.000	0.000	0.000
NA	0.000	0.000	0.000	0.000	0.000	0.000	0.444	1.000	0.615
Validation									
TUC-MI-B	0.511	0.382	0.386	0.577	0.509	0.541	0.737	0.862	0.794
ECNU_ICA-A	0.467	0.329	0.345	0.655	0.478	0.553	0.667	0.927	0.775
LQRZ-B	0.434	0.397	0.385	0.541	0.546	0.543	0.846	0.835	0.840
TUC-MI-A	0.461	0.322	0.330	0.542	0.463	0.500	0.721	0.872	0.789
LQRZ-A	0.468	0.344	0.355	0.636	0.495	0.557	0.696	0.920	0.793
ECNU_ICA-B	0.483	0.313	0.331	0.603	0.454	0.518	0.677	0.920	0.780
NICTA	0.485	0.297	0.324	0.649	0.398	0.493	0.597	0.931	0.727
Random	0.018	0.025	0.018	0.018	0.030	0.022	0.437	0.031	0.057
Majority	0.001	0.029	0.003	0.050	0.085	0.063	0.000	0.000	0.000
NA	0.000	0.000	0.000	0.000	0.000	0.000	0.409	1.000	0.580

(7th) exceed the heading-specific *F1* of 0.900 for 5, 6, 3, and 2 headings, respectively. These four methods have also the respective heading-specific *F1* of 0.807, 0.802, 0.813, and 0.749 in filtering out irrelevant information. In comparison, clinical IE has gradually improved to reach this heading-specific *F1* of at least 0.900 in 1995–2008 but none of these 170 reviewed studies focuses on nursing notes [33]. Instead, they report on processing chest and other types of radiography reports, discharge summaries, echocardiogram reports, and pathology reports.

As discussed in our 2015 Task 1a overview [9], *open data*, *open source code*, and *open evaluation results* are not only prerequisites for the basic scientific principle of the result reproducibility [34, 35] but they also increase return on public investment, encourage diversity of studies and opinion, enable the exploration of new topics and areas, and reinforce open scientific inquiry [36, 37]. Whilst this open movement in health sciences and informatics is progressing, particularly for source code [38] and evaluation results [39], its slowness in releasing data has significantly hindered method research, development, and adoption [40].

Evaluation labs have improved the situation [40, 41], but with some exceptions [42–44],[14] most open data are *deidentified* [14, 15] and/or *with use restriction* [45, 15]. The risk of identifiable components remaining in deidentified data and other difficulties in deidentification [46, 47] are against releasing verbal clinical documents or their transcriptions [48]. Moreover, on Australian clinical data, deidentification is to be avoided because under the *Australian Privacy Act*, it actually results in ***re***_identifiable data_, which introduces use restriction together with requirements to consent all data subjects (e.g., patients, their visitors, nurses, and other clinicians in the case of Australian nursing shift-change handover with nurses team meeting followed by a patient-site meeting) and obtain the the proper ethics approvals and research permissions [49, p. 27]. The use restrictions are further complicated for our case of the Australian nursing shift-change handover at the patient site, since non-reidentifiable real documents applicable to this case that allow the release and use without, for example, commercial restriction do *not exist*.

The significance of this study lies in its releases of our open synthetic clinical data, open source code, and open evaluation benchmarks to support innovation and decreasing barriers of method research and development. Due to the aforementioned difficulty of providing ethically-sound open data, we have compromised by providing synthetic data that closely matches the reality. We have validated this matching by employing and project-funding clinical experts to confirm the typicality and compare the synthetic documents with real documents, forms, and related processing results [10, 19, 11, 7, 8]. To the best of our knowledge, this is the first open dataset that that matches our case of the Australian nursing shift-change handover at the patient site, is not reidentifiable, and can be used without restriction. Moreover, to lower the entry barrier and encourage novelty, we have released both evaluation and processing code for IE together with its all intermediate and final outputs for the participants and other members of the clinical language processing community as an option to build on [8].

Acknowledgments

This shared task was partially supported by the CLEF Initiative and NICTA, which was supported by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Center of Excellence Program. We express our gratitude to Adj/Prof Leif Hanlen at NICTA for helping us to design the handover form and NICTA method — used as a baseline in Task 1 — to fill it out automatically and Ms Maricel Angel, Registered Nurse at NICTA, for helping us to create the Task 1 data sets, using the Protégé resource, which is supported by grant

[14] Synthetic clinical documents, written by clinicians about imaginary patients, have been used in the evaluation labs of the NII Test Collection for Information Retrieval Systems (NTCIR) for Japanese medical documents in 2013 (`http://mednlp.jp/medistj-en/`), 2014 (`http://mednlp.jp/ntcir11/`), and 2015 (`https://sites.google.com/site/mednlpdoc/` (last accessed on 1 July 2016).

GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. Last but not least, we gratefully acknowledge the participating teams' hard work. We thank them for their submissions and interest in the task.

References

1. Glaser, S. R., Zamanou, S., Hacker, K.: Measuring, interpreting organizational culture. Management Communication Quarterly (MCQ) 1(2), 173–198 (1987)
2. Tran, D. T., Johnson, M.: Classifying nursing errors in clinical management within an Australian hospital. International Nursing Review 57(4), 454–462 (2010)
3. Finlayson, S. G., LePendu, P., Shah, N. H.: Building the graph of medicine from millions of clinical narratives. Scientific Data 1, 140032 (2014)
4. Pothier, D., Monteiro, P., Mooktiar, M. Shaw, A.: Pilot study to show the loss of important data in nursing handover. British Journal of Nursing 14(20), 1090–1093 (2005).
5. Matic, J. Davidson, P., Salamonson, Y.: Review: Bringing patient safety to the forefront through structured computerisation during clinical handover. Journal of Clinical Nursing 20(1–2), 184–189 (2011)
6. Australian Commission on Safety and Quality in Healthcare (ACSQHC): Standard 6: Clinical handover. In: National Safety and Quality Health Standards, pp. 44–47. ACSQHC, Sydney, NSW, Australia (2012)
7. Suominen, H., Johnson, M., Zhou, L., Sanchez, P., Sirel, R., Basilakis, J., Hanlen, L., Estival, D., Dawson, L., Kelly, B.: Capturing patient information at nursing shift changes: Methodological evaluation of speech recognition and information extraction. Journal of the American Medical Informatics Association (JAMIA) 22(e1), e48–e66 (2015)
8. Suominen, H., Zhou, L., Hanlen, L, Ferraro, G.: Benchmarking clinical speech recognition and information extraction: New data, methods, and evaluations. JMIR Medical Informatics 3(2), e19 (2015)
9. Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., Jones, G.J.: Task 1a of the CLEF eHealth evaluation lab 2015: Clinical speech recognition. In: CLEF 2015 Online Working Notes, CEUR-WS (2015)
10. Dawson, L., Johnson, M., Suominen, H., Basilakis, J., Sanchez, P., Kelly, B., Hanlen, L.: A usability framework for speech recognition technologies in clinical handover: A preimplementation study. Journal of Medical Systems 38(6), 1–9 (2014)
11. Johnson M, Sanchez P, Suominen H, Basilakis J, Dawson L, Kelly B, Hanlen L.: Comparing nursing handover and documentation: Forming one set of patient information. International Nursing Review 61(1), 73–81 (2014)
12. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., NÃľvÃľol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. In: CLEF 2015 — 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS) 9283, pp. 429–443. Springer, Berlin Heidelberg, Germany (2015)
13. Kelly, L., Goeuriot L., Suominen, H., N/'ev/'eol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: CLEF 2016 — 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer, Berlin Heidelberg, Germany (2016)

14. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.): Information Access Evaluation. Multilinguality, Multimodality, and Visualization, LNCC, vol. 8138, pp. 212–231. SpringerVerlag, Berlin Heidelberg, Germany (2013)
15. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D. L., Velupillai, S., Chapman, W. W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: CLEF 2014 — 5th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS) 8685, pp. 172–191. Springer, Berlin Heidelberg, Germany (2014)
16. Poissant, L., Pereira, J., Tamblyn, R., Kawasumi, Y.: The impact of electronic health records on time efficiency on physicians and nurses: A systematic review. Journal of the American Medical Informatics Association (JAMIA) 12(5), 505–516 (2005)
17. Hakes, B., Whittington, J.: Assessing the impact of an electronic medical record on nurse documentation time. Journal of Critical Care 26(4), 234–241 (2008)
18. Banner, L.., Olney, C.: Automated clinical documentation: Does it allow nurses more time for patient care? Computers, Informatics, Nursing (CIN) 27(2), 75–81 (2009)
19. Johnson, M., Lapkin, S., Long, V., Sanchez, P., Suominen, H., Basilakis, J., Dawson, L.: A systematic review of speech recognition technology in health care. BMC Medical Informatics and Decision Making 14, 94 (2014)
20. Williams, J. R.: Guidelines for the use of multimedia in instruction. In: Proceedings of the Human Factors, Ergonomics Society 42nd Annual Meeting, pp. 1447–1451 (1998)
21. Ayres, R. U., Martinás, K.: 120 wpm for very skilled typist. In: On the Reappraisal of Microeconomics: Economic Growth and Change in a Material World, p. 41. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, USA (2005)
22. Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J., Zhai, C.: Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. SIGIR Forum 37(1), 31–47 (2003)
23. Ogren, P. V.: Knowtator: A Protégé plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 273–275. Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
24. Lafferty, J. D., McCallum, A., Pereira, F. C. N.: Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, ICML 2001. Morgan Kaufmann, Burlington, MA, USA 2001.
25. Zhou, L., Suominen, H., Hanlen, L.: Evaluation data and benchmarks for cascaded speech recognition and entity extraction. In: ACM Multimedia 2015 Workshop on Speech, Language and Audio in Multimedia, pp. 15–18. Association for Computing Machinery, New York, NY, USA (2015)

26. Zhou, L., Suominen, H.: Information extraction to improve standard compliance: The case of clinical handover. In: AI 2015: — Advances in Artificial Intelligence, Lecture Notes in Computer Science (LNCC) 9457, pp. 644–649. Springer, Berlin Heidelberg, Germany (2015)
27. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Stroudsburg, PA, USA (2014)
28. MetaMap — A tool for recognizing UMLS concepts in text. `http://metamap.nlm.nih.gov/`, last accessed on 1 July 2016 (2012)
29. Ontoserver. `http://ontoserver.csiro.au:8080/`, last accessed on 1 July 2016 (2013)
30. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83 (1945)
31. Shapiro, S. S., Wilk, M. B.: An analysis of variance test for normality (complete samples). Biometrika 52(3âĂŞ4), 591–611 (1965)
32. Razali, N., Wah, Y. B.: Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling., Analytics 2(1), 21–33 (2011)
33. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., Hurdle, J. F.: Extracting information from textual documents in the electronic health record: A review of recent research. The International Medical Informatics Association (IMIA) Yearbook of Medical Informatics, 128–144 (2008)
34. Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCunn, Y., Müller, K.-R., Pereira, F., Rasmussen, C. E., Rätsch, G., Schölkopf, B., Smola, A., Vincent, P., Weston, J., Williamson, R. C.: The need for open source software in machine learning. Journal of Machine Learning 8, 2443–2466 (2007)
35. Pedersen, T.: Empiricism is not a matter of faith. Computational Linguistics 34(3), 465–470 (2008)
36. Organisation for Economic Development (OECD): OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD, Danvers, MA, USA (2007)
37. Jisc: The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. `http://www.jisc.ac.uk/whatwedo/programmes/di_directions/strategicdirections/textmining.aspx`, last accessed on 1 July 2016 (2012)
38. Estrin, D. and Sim, I.: Open mHealth architecture: An engine for health care innovation. Science 330(6005), 759–760 (2010)
39. Dunn, A. G., Day, R. O., Mandl, K. D., Coiera, E.: Learning from hackers: open-source clinical trials. Science Translational Medicine 4(132), 132cm5 (2012)
40. Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., Uzuner, Ö: Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. Editorial. Journal of the American Medical Informatics Association: JAMIA 18(5), 540–543 (2011)
41. Huang, C.-C. and Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics 17(1) 132–144 (2016)
42. Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., Aramaki, E.: Overview of the NTCIR-10 MedNLP task. In: Proceedings of the 10th NTCIR Conference, pp. 696–701. NTCIR, Tokyo, Japan (2013)

43. Aramaki, E., Morita, M., Kano, Y., Ohkuma, T.: Overview of the NTCIR-11 MedNLP-2 task. In: Proceedings of the 11th NTCIR Conference, pp. 147–154. NTCIR, Tokyo, Japan (2014)
44. Aramaki, E., Morita, M., Kano, Y., Ohkuma, T.: Overview of the NTCIR-12 MedNLPDoc task. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, pp. 71–75. NTCIR, Tokyo, Japan (2015)
45. Neamatullah,, I., Douglass, M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., Clifford, G. D.: Automated deidentification of free-text medical records. BMC Medical Informatics and Decision Making 8, 32 (2008)
46. Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., Salanterä, S.: Applying language technology to nursing documents: Pros and cons with a focus on ethics. International Journal of Medical Informatics 76(S2), S293–S301 (2007)
47. Carrell, D.Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., Hirschman, L.: Hiding in plain sight: Use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association: JAMIA 20(2), 342–348 (2013)
48. Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., Altman, D. G.: Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. British Medical Journal (BMC) 340, c181 and Trials 11, 9 (2010)
49. National Health and Medical Research Council, Australian Research Council and Australian Vice-Chancellors' Committee: National Statement on Ethical Conduct in Human Research. National Health Medical Research Council and Australian Vice-ChancellorsâĂŹ Committee, Canberra, ACT, Australia (2007 updated 2014)