Task 1a of the CLEF eHealth Evaluation Lab 2015

Clinical Speech Recognition

Hanna Suominen[1*], Leif Hanlen[2], Lorraine Goeuriot[3], Liadh Kelly[4], Gareth J F Jones[5]

[1] NICTA, The Australian National University (ANU), University of Canberra (UC), and University of Turku, Canberra, ACT, Australia, hanna.suominen@nicta.com.au
[2] NICTA, ANU, and UC, Canberra, ACT, Australia, leif.hanlen@nicta.com.au
[3] Université Grenoble Alpes, Grenoble, France, lorraine.goeuriot@imag.fr
[4] Trinity College Dublin, Dublin, Ireland, liadh.kelly@scss.tcd.ie
[5] Dublin City University, Dublin, Ireland, Gareth.Jones@computing.dcu.ie
[*] Corresponding author

Abstract. Best practice for clinical handover and its documentation recommends standardized, structured, and synchronous processes with patient involvement. Cascaded speech recognition (SR) and information extraction could support their compliance and release clinicians' time from writing documents to patient interaction and education. However, high requirements for processing correctness evoke methodological challenges. First, multiple people speak clinical jargon in the presence of background noise with limited possibilities for SR personalization. Second, errors multiply in cascading and hence, SR 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 in 2015. The task released 100 synthetic handover documents for training and another 100 documents for testing in both verbal and written formats. It attracted 48 team registrations, 21 email confirmations, and four method submissions by two teams. The submissions were compared against a leading commercial SR engine and simple majority baseline. Although this engine performed significantly better than any submission [i.e., 38.5 vs. 52.8 test error percentage of the best submission with the Wilcoxon signed-rank test value of 302.5 ($p < 10^{-12}$)], the releases of data, tools, and evaluations contribute to the body of knowledge on the task difficulty and method suitability.

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

Contributor Statement: HS, LH, and GJFJ designed the task and its evaluation methodology. HS developed the dataset and together with LH, led the task as a part of the CLEFeHealth2015 evaluation lab, chaired by LG and LK. HS drafted the paper and after this all authors revised and approved it.

1 Introduction

Fluent *information flow*, defined as channels, contact, communication, or links to pertinent people [1], is critical in healthcare in general and in particular in *clinical handover* (aka *handoff*), when a clinician or group of clinicians is transferring professional responsibility and accountability, for example, at shift change of nurses [2]. This *shift-change nursing handover* is a form of clinical narrative where only a small part of the flow is documented in writing [3]. Best practice recommends standardized, structured, and synchronous processes for handover and its information documentation not only in the presence but also in active involvement of the patients, and where relevant, their next-of-kin [4].[1] However, failures in information flow 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]. Only after a couple of shift changes, anything from two-thirds to all *verbal* handover information is lost or, even worse, transferred incorrectly if *not* documented *electronically in writing* [5, 6].

In order to support compliance with these processes, cascaded *speech recognition* (SR) with *information extraction* (IE) has been studied in 2015 [7, 8]. As justified empirically in clinical settings in 2014, the cascade pre-fills a structured handover form for a clinician to proof and sign off [9, 10]. Based on the aforementioned rate of information loss, the approach of the nurse who is handing over proofing and signing off the document draft him/herself any time before the shift ends (but preferably immediately after the handover) can decrease the loss to 0–13 per cent.

This novel application evokes fruitful challenges for method research and development, and consequently, its first part (i.e., clinical SR) was chosen as the *Task 1a* of the *eHealth Evaluation Lab* by the *Conference and Labs of the Evaluation Forum* (CLEF) in 2015 [11].[2] First, clinical characteristics complicate SR. This derives from a large number of nursing staff 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 and, because of the severe implications that they may have in clinical decision-making, the cascade correctness needs to be carefully evaluated as meeting the clinical requirements.

The task aligns with the CLEFeHealth usage scenario of easing patients, their next-of-kin, and other *laypersons* in understanding and accessing *electronic health* (eHealth) information [12, 13]. Namely, the application could release a substantial amount of nurses' time from documentation to, for example, longer discussions about the findings, care plans, and consumer-friendly resources for further information with

[1] Also the *World Health Organisation* (WHO) provides similar guidance as a mechanism to contribute to safety and quality in healthcare at `http://www.who.int/patientsafety/research/methods_measures/human_factors/organizational_tools/en/` (all websites of this paper were accessible on 25 May 2015)

[2] `https://sites.google.com/site/clefehealth2015/`

the patients, and where relevant, their next-of-kin. Documenting every event in healthcare, as required by law, can take nearly sixty per cent 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) [14–16]. For example, every year within the *Organisation for Economic Co-operation and Development* (OECD), on average seven physician consultations and 0.2 hospital discharges take place per capita.[3] SR writes a document draft from a tenth to three-quarters of the time it takes to transcribe this by hand, whilst the clinician's proofing time is about the same in both cases [17]. The speech-recognized draft for a minute of verbal handover (with 160 words, corresponding to the range that people comfortably hear and vocalize words [18]) is available only 20 seconds after finishing the handover with a real-time engine that recognizes at least as many words per minute as a very skilled typist (i.e., 120 [19]). Cascading this with content structuring through IE can bring further efficiency gains by easing finding information and making this content available for computerized decision-making and surveillance in healthcare [20].

2 Materials and Methods

The *hold-out method* was used in *performance evaluation* of the task. Task materials consisted of a *training set* of 100 synthetic patient cases and an *independent set* of another 100 synthetic patient cases *for testing*. Given the training set, the task was to minimize the number of incorrectly recognized words on the test set (i.e., on the held-out set). Performance of the submitted methods and two baseline methods was compared statistically.

2.1 Dataset for Training

The dataset called *NICTA Synthetic Nursing Handover Data* was used in this task for method development and training [8].[4] This set of 100 synthetic patient cases was developed for SR and IE related to nursing shift-change handover in 2012–2014. Each case consisted of a *patient profile*; a written, free-form text paragraph (i.e., the *written handover document*) to be used as a reference standard in SR; and its spoken (i.e., the *verbal handover document*) and speech-recognized (i.e., *speech-recognized documents* with respect to six vocabularies) counterparts. The dataset was released on the task page on 15 November 2014.

First, the first author of this paper (Adj/Prof in machine learning and communication for health computing) generated 100 synthetic patient profiles, using common user profile generation techniques [21]. With an aim for balance in patient types, she created profiles for 25 *cardiovascular*, 25 *neurological*, 25 *renal*, and 25 *respiratory*

[3] Derived from *OECD.StatsExtracts* (`http://stats.oecd.org/`) for 2009 (i.e., the most recent year that has almost all data available)

[4] `http://www.nicta.com.au/nicta-synthetic-nursing-handover-open-data-software-and-demonstrations/`

patients of an imaginary *medical ward for adults in Australia*. These patient types were chosen because they represent the most common chronic diseases and national priority areas [22]. The reason for patient admission was always an acute condition, but some patients had also chronic diseases. Some patients were recently admitted to the ward, some had been there for some days already, and some were almost ready to be discharged after a shorter or longer inpatient period. Each profile was saved as a DOCX file and contained a stock photo from a royalty-free gallery, name, age, admission story, in-patient time, and familiarity to the handover nurses.

Second, the first author supervised a *registered nurse* (RN) in creating the written handover documents for these 100 profiles. The RN had over twelve years' experience in clinical nursing. Australian English was her second language and she was originally from the Philippines. She was guided to imagine herself working in the medical ward and delivering verbal shift-change handovers to another nurse by the patient's bedside as if she was talking. All handover information was to be given as a 100–300-word monologue, using normal wordings. The resulting realistic but fully imaginary handovers were saved as TXT files.

Third, the first author supervised the RN in creating the verbal handover documents by reading the written handover documents out loud as the nurse giving the handover. She was guided to record in a quiet office environment, try to speak as naturally as possible, avoid sounding like reading text, and repeat the take until she was satisfied with the outcome. The *Olympus WS-760M digital recorder* [purchased for 269.00 AUD (191 €) in October 2011, weight of 51 g, dimensions of 98.5 mm × 40.0 mm × 11.0 mm] and *Olympus ME52W noise-canceling lapel-microphone* [purchased for 15.79 AUD (11 €) in October 2011, weight of 4 g (+ 11 g for an optional cable and clip), dimensions of 35 mm (+ a 15 mm plug) × 13 mm × 13 mm] were used, because they were previously shown to produce superior word correctness in SR [7].[5, 6] Each document was saved as a WMA file and then converted from stereo to mono tracks and exported as WAV files on *Audacity 2.0.3 for Mac*.[7]

Fourth, the first author used *Dragon Medical 11.0* for SR.[8] This clinical engine was chosen because it included an option for Australian English. It was first initialized with not only this accent but also to the RN's age of 22–54 years and recording of The Final Odyssey [DOCX file of 3,893 words in writing and WMA file of 29 minutes 22 seconds as speech (4 minutes needed)] using the aforementioned recorder and microphone. Also these training/personalization/adaption files were released. Six Dragon vocabularies (i.e., *general* as the most generic clinical vocabulary, *medical* because of the medical ward, *nursing* because of the nursing handover, *cardiology* because of the cardiac patients, *neurology* because of the neurological patients, and *pulmonary disease* because of the respiratory patients) were compared although the nursing vocabu-

[5] http://www.olympus.co.uk/site/en/archived_products/audio/audio_recording_1/ws_760m/index.pdf

[6] https://shop.olympus.eu/UK-en/microphones/olympus/me52w-mini-mono-microphone-p-239.htm

[7] http://sourceforge.net/projects/audacity/

[8] http://www.nuance.com/products/dragon-medical-practice-edition/index.htm

lary shown to produce the best results in SR [7, 8]. Each speech-recognized document was saved as a TXT file.

The data release with the requirement to cite [8] was approved at NICTA and the RN was consented in writing. The license of the verbal, free-form text documents (i.e., WMA and WAV files) was *Creative Commons - Attribution Alone - Non-commercial - No Derivative Works* for the purposes of testing SR and language processing algorithms.[9] The remaining documents (i.e., DOCX and TXT files) were licensed under *Creative Commons – Attribution Alone*.[10]

2.2 An Independent Dataset for Testing

The training set was supplemented with *an independent dataset for testing*. This additional set of 100 synthetic patient cases was developed in 2015. Each case consists of (1) a patient profile, (2) a written handover document, (3) a verbal handover document, and (4) a speech-recognized document with respect to the nursing vocabulary. Its subset of documents (3) and (4) was released on the task page on 23 April 2015; the organizers did not release the profiles in order to avoid their contents to be used an a processing input, and they held the written handover documents out as a blind set for anyone but the first author and RN to ensure independent training and testing. The set was created the same way as the training set except that the profile photos were reused, software was updated to *Audacity 2.1.3 for Mac* with *ffmpeg-mac-2.2.2*,[11] and only the nursing vocabulary was chosen for SR.

The data release was approved at NICTA and the RN was consented in writing. The licensing constraints are the same as before; however, we ask to cite this task overview for the data release.

2.3 Submission to Performance Evaluation

The participants needed to submit their processing results by 1 May 2015 using the *Easy Chair System of the lab*.[12] Submissions that developed the *SR engine itself* were evaluated separately from those that studied *post-processing methods for the speech-recognized text*. Also a separate *submission category* was assigned to solutions based on *both SR and text post-processing*.

Only fully automated methods were allowed, that is, human-in-the-loop means were not permitted. Each participant was allowed to submit up to two methods/parameterizations/compilations (referred to as a *method* from now on) to the first category and up to two methods to the second category. If addressing both these categories, the participant was asked to submit all possible combinations of these methods as their third category submission (i.e., up to $2 \times 2 = 4$ files).

---

[9] `http://creativecommons.org/licenses/by-nc-nd/4.0/`

[10] `http://creativecommons.org/licenses/by/4.0/`

[11] `https://www.ffmpeg.org/`

[12] `https://easychair.org/conferences/?conf=clefehealth2015resul` with the professional license

In addition to the submission category, the submissions consisted of the following elements: *team name* and *description*; *address of correspondence*; *author(s)*; at least three *method keywords*;[13] *method description* (max 100 words per method); and *processing outputs* for each method on the 100 training and 100 test documents.

2.4 Methods and Measures in Performance Evaluation

We challenged the participants to minimize the number of incorrectly recognized words on the independent test set. This correctness was evaluated on the entire test set using the primary measure of the percentage of incorrect words [aka the *error rate percentage* (*E*)] as defined by the *Speech Recognition Scoring Toolkit (SCTK), 2.4.0 without* punctuation as a differentiating feature.[14] This measure sums up the *percentages of substituted* (*S*), *deleted* (*D*), and *inserted* (*I*) words (i.e., $E = S + D + I$) and consequently, the smaller the value of *E*, the better the performance. To illustrate these error types, speech-recognizing *your word* as *you are* had the substitution (*your*, *you*), insertion *are*, and deletion *word*.

As secondary measures, we reported the *percentage of correctly detected words* (*C*) on the entire test set together with the breakdown of *E* to *D*, *I*, and *S*. We also documented the raw numbers of correct (n_C), substituted (n_S), deleted (n_D), and inserted words (n_I). Notice that $C + S + D = 100$ and $n_C + n_S + n_D$ is the number of words in the reference standard.

To provide more details on performance differences across the individual handover documents, we also computed the error rate percentage e_i in individual documents $d_i \in \{d_1, d_2, d_3, \ldots, d_{100}\}$ of the test set. Then, we summarized these values through their *minimum* (min), *maximum* (max), *mean*, *median*, and *standard deviation* (SD).

Finally, instead of evaluating this generalization capability of the method to unseen data, we assessed the resubstitution performance on the training set; a method that does not perform well even on its training set is poor, but excellence on training set may indicate over-fit, leading to issues in the generalizability.

2.5 Baselines Methods in Performance Evaluation

We used two baseline systems in the task, namely *Dragon* and *Majority*. The Dragon baseline was based on Dragon Medical 11.0 with the nursing vocabulary and initialization to the RN, recorder, and microphone. This commercial system included substantial but closed domain dictionaries, had a substantial license fee per person [purchased for 1,600.82 AUD (1,139 €) in January 2013], and was limited to the Microsoft Windows operating system. Notice that the 200 synthetic handover cases were *not* used to train the Dragon baseline. The Majority baseline assumed that the right number of words is detected (i.e., the number of test words originating from the refer-

[13] preferably Medical Subject Headings (`http://www.nlm.nih.gov/mesh/MBrowser.html`) or Association for Computing Machinery classes (`http://www.acm.org/about/class/ccs98-html`)

[14] `http://www.itl.nist.gov/iad/mig/tools/`

ence standard) and recognized each word as the most common training word (i.e., *and*) with correct capitalization (i.e., *and* for *and*, *And* for *And*, and so forth).

To supplement the usage guidelines of SCTK,[15] we provided the participants some helpful tips (Appendix 1): we released an example script for removing punctuation and formatting text files; a formatted reference file and Dragon baseline for the training set; overall and document-specific evaluation results for this file pair; and commands to perform these evaluations and ensure the correct installation of SCTK.

2.6 Statistical Significance Testing

Statistical differences between the error rate percentages of the two baselines and participant submissions were evaluated using the *Wilcoxon signed-rank test* (*W*) [23]. This test was chosen as an alternative to the paired *t*-test, because the *Shapiro-Wilk test* [24, 25] with the significance level of 0.005 indicated that the error rate percentages e_i for the sample of the 100 test documents were not normally distributed (e.g., *p* values of 0.018 and 0.225 for the Dragon and Majority baselines, respectively).

After ranking the baselines and submissions based on their error rate percentage on the entire dataset for testing, *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. All statistical tests were computed using *R 3.2.0*.[16]

3 Results

The task released in both verbal and written formats the total of 200 synthetic clinical documents that can be used for studies on nursing documentation and informatics. It attracted nearly 50 team registrations with about half of them confirming their participation through email. Two teams submitted two SR methods each. Although *no* method performed as well as the Dragon baseline, the task contributed to the body of knowledge on the task difficulty and method suitability.

3.1 Data Release

The task released a training set of 100 documents (Fig. 1) on 27 October 2014; an independent test set of 100 documents on 23 April 2015; and reference standard for the test documents together with processing results of the Dragon baseline and sub-

Ken harris, bed three, 71 yrs old under Dr Gregor, came in with arrhythmia. He complained of chest pain this am and ECG was done and was reviewed by the team. He was given some

[15] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/options.htm#option_r_name_0

[16] http://www.r-project.org/

anginine and morphine for the pain. Still tachycardic and new meds have been ordered in the medchart. still for pulse checks for one full minute. Still awaiting echo this afternoon. His BP is just normal though he is scoring MEWS of 3 for the tachycardia. He is still for monitoring.

Dragon baseline: *Own now on bed 3 he is then Harry 70 is 71 years old under Dr Greco he came in with arrhythmia he complained of chest pain this morning in ECG was done and reviewed by the team he was given some and leaning in morphine for the pain in she is still tachycardic in new meds have been ordered in the bedtime is still 4 hours checks for one full minute are still waiting for echocardiogram this afternoon he is BP is just normal though he is scarring meals of 3 for the tachycardia larger otherwise he still for more new taurine*

Fig. 1. Speech-recognized training document

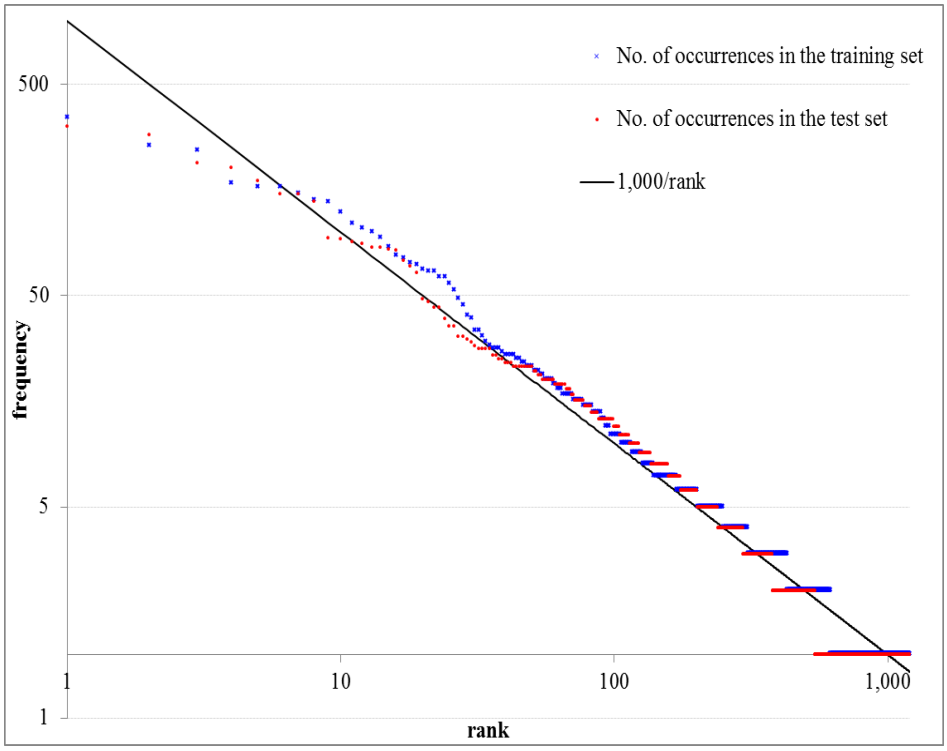

Fig. 2. Both on the training set and test set, the word frequency (*y*-axis on a logarithmic scale) was inversely proportional to the rank of the word commonness (*x*-axis on a logarithmic scale). Capitalization was *not* considered as a differentiating feature.

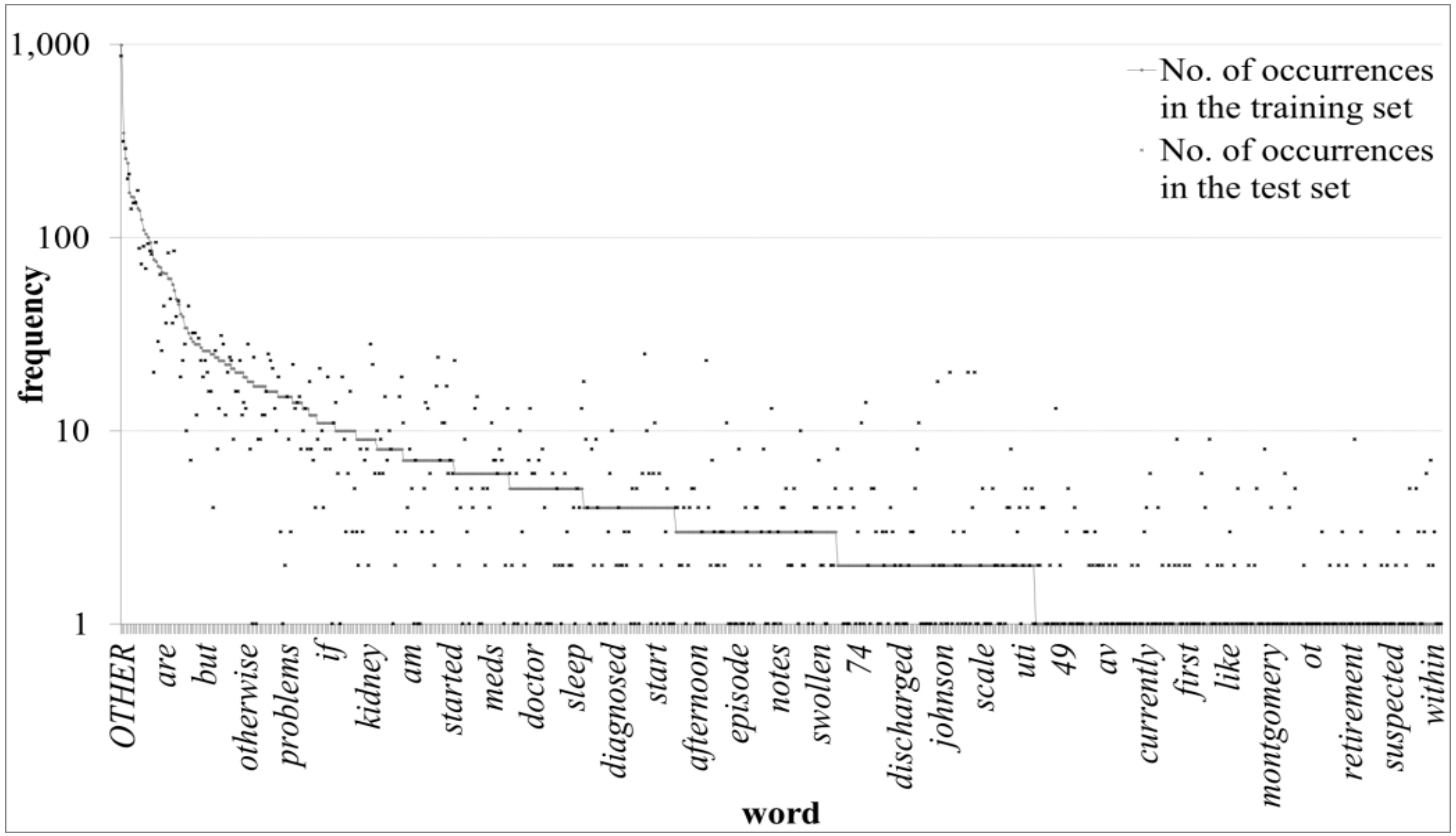

Fig. 3. Distribution of unique words in the training set and test set *without* capitalization as a differentiating feature. *OTHER* refers to words that did *not* occur in both sets. The *y*-axis uses a logarithmic scale and the *x*-axis is in decreasing order of word frequency on the training set.

missions later in 2015.[17] Errors that the Dragon baseline made on training documents ten or more times included substituting *years* for *yrs* (*n* = 48), *in* with *and* (*n* = 22), *one* with *1* (*n* = 17), *alos* with *obs* (*n* = 12), and *to* with *2* (*n* = 12); deleting *is* (20), *are* (13), and *and* (11); and inserting *and* (210), *is* (136), *in* (106), *she* (71), *are* (58), *all* (45), *arm* (44), *for* (43), *the* (37), *he* (35), *that* (34), *a* (27), *her* (19), *eats* (15), *on* (15), *also* (14), *am* (12), *does* (11), *bed* (10), *s* (10), and *to* (10) [26].

The training set had 7,277 words and 1,304 (1,377) of them were unique *without* (with) capitalization as a differentiating feature. For the test set, these numbers were 6,818, 1,279, and 1,323, respectively. Although both sets followed the *Zipf's law* [27] (Fig. 2), and were thereby typical language samples, their vocabularies shared only 645 (738) unique words *without* (with) capitalization as a differentiating feature (Fig. 3). Consequently, the sets can be seen as fairly independent, as intended. The 10 common words *without* capitalization as a differentiating feature were *and* (347 occurrences in the training asset and 315 in the test set), *is* (256, 288), *he* (243, 201), *in* (170, 212), *for* (163, 140), *with* (162, 151), *she* (151, 152), *on* (141, 175), *the* (138, 88), and *to* (124, 73).

3.2 Community Interest and Participation

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 nursing expertise.

By 30 April 2015, 48 people had registered their interest in the task through the *CLEF 2015 registration system*,[18] and 21 of these team leaders had emailed to confirm their participation. From its opening on 27 October 2014 to this date, the task *discussion forum* gained five other members than the organizers.[19]

By 1 May 2015, two teams submitted four methods. The first team, called *TUC_MI/MC*, was from the *Technische Universität Chemnitz* (TUC) in Germany. Its members were two researchers from the field of computer science, supervised by two TUC professors. They followed an interdisciplinary approach where one part brought the expertise from the field of speech processing to develop strategies for web-based language model adaptation. The other one came from the field of information retrieval to choose and develop methods for selecting and processing web resources to build a thematically coherent adaptation corpus. The second team, called *UC*, was from the *University of Canberra* in the Australian Capital Territory. It consisted of two PhD

[17] http://www.nicta.com.au/nicta-synthetic-nursing-handover-open-data-software-and-demonstrations/.

[18] http://clef2015-labs-registration.dei.unipd.it/

[19] https://groups.google.com/forum/#!forum/clefehealth2015-task-1a-speech-recognition

students and three Professors from multi-disciplinary backgrounds, including clinical, public health, machine learning, and software engineering, working in collaboration.[20]

TUC_MI/MC submitted two SR methods. Their approach assumed each document having its own context and hence suggested adapting SR for each document separately. They used a two-pass decoding strategy: First, a verbal document was speech recognized. Then, keywords of the utterances were extracted and used as queries in order to retrieve web resources as adaptation data to build a document-specific dictionary and language model with the interpolation weights of 0.8 and 0.9 for *TUC_MI/MC.1* and *TUC_MI/MC.2*, respectively. Finally, re-decoding of the same document was performed using the adapted dictionary and language model.

Also *UC* submitted two SR methods. *UC.1* was based on acoustic modeling of speech using Hidden Markov Models (HMM). The verbal documents were pre-processed, including filtering, word level segmentation, and Mel Frequency Cepstral feature extraction, and HMM models were built for the training data. As there were no repetitions of data from different sessions, bagging and bootstrapping of training data were used. *UC.2* combined language and acoustic models using the CMU Sphinx open source toolkit for SR. A custom dictionary and language model was developed for the speaker of the training set, because none of existing dictionary and language models was suitable for her accent. Unfortunately, the organizers had to reject this second method as even after an update request and deadline extension to 10 May 2015, this submission failed to meet the evaluation criteria for the format and completeness.

3.3 Performance Benchmarks

The Dragon baseline clearly had the best performance (i.e., $E = 38.5$) on the independent set for testing, followed by the TUC_MI/MC.2 ($E = 52.8$), TUC_MI/MC.1 ($E = 52.3$), UC.1 ($E = 93.1$), and the Majority baseline ($E = 95.4$) (Table 1). The resubstitution performance of the first three methods was approximately the same (i.e., from $E = 55.0 \pm 0.9$), but last two methods had nearly 100 per cent error also on the training set (Table 2).

The performance of the Dragon baseline on the test set was significantly better than that of the second-best method (i.e., TUC_MI/MC.2, $W = 302.5$, $p < 10^{-12}$). However, this rank-2 method was *not* significantly better than the third-best method (i.e., TUC_MI/MC.1), but this rank-3 method was significantly better than the fourth-best method (i.e., UC.1, $W = 0$, $p < 10^{-15}$). Finally, the performance of the lowest-ranked method (i.e., the Majority baseline) was significantly worse than that of this rank-4 method ($W = 1{,}791.5$, $p < 0.05$).

[20] Including Prof LH, task co-leader, as an advisor who encouraged participation without any engagement in the team's method experimentation and development. As noted in Section 2.2, he did not develop the test set nor had an access to it before other participants.

Table 1. Performance of the baselines and submissions on the 100 test documents

	Dragon	TUC_MI/MC.2	TUC_MI/MC.1	UC.1	Majority
C	73.1	54.3	53.7	17.0	4.6
S	22.6	36.6	36.7	49.3	95.4
D	4.3	9.1	9.6	33.7	0.0
I	11.6	6.6	6.5	10.1	0.0
E	38.5	52.3	52.8	93.1	95.4
n_C	4,984	3,703	3,660	1,159	315
n_S	1,539	2,493	2,503	3,359	6,503
n_D	295	622	655	2,300	0
n_I	792	451	443	687	0
n_E	2,626	3,566	3,601	6,346	6,503
min(e_i)	20.7	26.2	26.2	65.8	89.7
max(e_i)	59.1	92.0	92.0	134.5	100.0
mean(e_i)	39.5	52.1	52.6	93.3	95.4
median(e_i)	39.1	51.6	51.7	93.8	95.6
SD(e_i)	9.8	12.9	13.1	10.4	2.2

Table 2. Performance of the baselines and submissions on the 100 training documents

	Dragon	TUC_MI/MC.2	TUC_MI/MC.1	UC.1	Majority
C	72.3	65.6	64.6	9.9	4.8
S	24.1	27.9	28.6	31.8	95.2
D	3.6	6.6	6.8	58.3	0.0
I	28.2	19.6	19.5	5.1	0.0
E	55.9	54.1	54.9	95.2	95.2
n_C	5,260	4,771	4,701	723	347
n_S	1,757	2,027	2,083	2,314	6,930
n_D	260	479	493	4,240	0
n_I	2,049	1,423	1,417	370	0
n_E	4,066	3,929	3,993	6,924	6,930

4 Discussion

We conclude the paper by comparing the results with prior work, validating the released data, and discussing the significance of this study.

Comparison with Prior Work

SR at its best can achieve an impressive *C* of 90–99 with only 30–60 minutes of personalization or adaptation to a given clinician's speech [17]. This SR correctness is supported by studies on mainly North-American male physicians speaking medical documents. For studio recordings of a Spanish-accented Australian female nurse, native Australian female nursing professional, and native Australian male physician

speaking nursing handover simulations, (*C*, *E*) pairs of Dragon Medical 11.0 are (62, 40), (64, 39), and (71, 32), respectively [7]. These numbers are very similar to those for the Dragon baseline on the training set [i.e., (72, 56)] and test set [i.e., (73, 39)]. Differences between commercial engines (i.e., *IBM ViaVoice 98, General Medicine* with $C = 92 \pm 1$, *L&H Voice Xpress for Medicine 1.2, General Medicine* with $C = 86 \pm 1$, and *Dragon Medical 3.0* with *C* from 85 to 86) are not drastic [28]. To compare this automation with the upper baseline of human performance, each clinical document has 0.4 errors on average if transcribing by hand whilst for a speech-recognized document, this number is 6.7 [29].

We have studied correcting SR errors through post-processing in [26]. This approach is unsupervised and applies phonetic similarity to substituted words. Its evaluation on the 100 training documents gives promising results; in 15 per cent of all 1,187 unique substitutions by the Dragon baseline, the speech-recognized word sounds exactly the same as its reference word and 23 per cent of them are at least 75 per cent similar.

Data Validation

A basic scientific principle of the *reproducibility of the results* relies on availability of *open data*, *open source code*, and *open evaluation results* [30, 31]. Access to research data also increases the returns from public investment in this area; encourages diversity of studies and opinion; enables the exploration of new topics and areas; and reinforces *open scientific inquiry* [32]. Whilst this *open movement* in health sciences and informatics is progressing, particularly for source code [33] and evaluation results from clinical trials [34], its slowness in releasing data has significantly hindered method research, development, and adoption [35]. Evaluation labs have improved the situation [35, 36], but with some exceptions [37, 38],[21] most open data are *de-identified* [12, 13] and/or *with use restriction* [39, 13].[22]

However, data de-identification on text documents is fraught with difficulties [40] and the resulting data may still have some identifiable components [41]. Consequently, the minimum standard of clinical data de-identification recommend *against* releasing verbal clinical documents or their transcriptions [42] – that is, precisely our open data. Furthermore, de-identification is to be *avoided* on Australian clinical data, because under Australian privacy law, it actually results in *re-identifiable data*, which must have restricted use, appropriate ethical use, and approval from 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 meet-

[21] Synthetic clinical documents 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/`) and 2014 (`http://mednlp.jp/ntcir11/`).

[22] These clinical data originating from US healthcare services accessible to registered users on a password-protected Internet site (i.e., *PhysioNetWorks* at `https://physionet.org/works/`) after manual authorization and approval of the data access and use policies (e.g., for lab-participation or scientific purposes only).

ing [43, p. 27]. These use restrictions are even more complicated for the Australian handover, as real documents that are not re-identifiable, apply to our patient-site case, and allow releasing and use without, for example, commercial restriction do *not exist*.[23]

Due to the lack of existing text corpora, that match the Australian clinical setting, and due to the difficulty of providing ethically-sound open data, we have compromised by providing synthetic data that closely matches the real data typically found in a nursing shift-change. We have validated this matching by employing and project-funding clinical experts to confirm the typicality and compare the synthetic documents with real data and related processing results [9, 17, 10, 7, 8].

Significance

The significance of our open synthetic clinical data lies in supporting innovation and decreasing barriers of method research and development. In particular for new participants in SR, IE, and other automated generation or analysis of text documents, the aforementioned barrier of data access costs money and time. The entry and transaction are even more expensive, but given the required expertise in the field, only the latter cost can be decreased substantially by simplified licensing through open data movement [44]. This also resolves the barrier of data access.

Acknowledgements

This shared task was partially supported by the CLEF Initiative and NICTA, funded by the Australian Government through the Department of Communications and the Australian Research Council through the Information and Communications Technology (ICT) Centre of Excellence Program. We express our gratitude to Maricel Angel, Registered Nurse at NICTA, for helping us to create this dataset. 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)

[23] The written documents captured in existing clinical systems in Australia are typically entered by a clerk later in the shift and hence are not nursing handover transcripts. In other works [7, 9, 10], we placed microphones on nurses to evaluate real (de-identified) data. Consequently, they cannot be released.

4. 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)
5. 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).
6. 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)
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. 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 pre-implementation study. Journal of Medical Systems 38(6), 1–9 (2014)
10. 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 2014 61(1), 73–81 (2014)
11. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, 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). Springer, Berlin Heidelberg, Germany (2015)
12. 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. Springer-Verlag, Berlin Heidelberg, Germany (2013)
13. 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: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.): Information Access Evaluation. Multilinguality, Multimodality, and Visualization, LNCC, vol. 8685, pp. 172–191. Springer-Verlag, Berlin Heidelberg, Germany (2014)
14. Poissant, L., Pereira, J., Tamblyn, R., Kawasumi, Y.: The impact of electronic health records on time efficiency on physicians and nurses: A systematic review. JAMIA 12(5), 505–516 (2005)
15. 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)
16. 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)
17. 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., Decision Making 14, 94 (2014)
18. 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)

19. 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 & Northampton, MA, USA (2005)
20. 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., 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)
21. Kuniavsky, M.: Observing the User Experience: A Practitioner's Guide to User Research. Morgan Kaufmann Publishers, San Fransisco, CA, USA (2003)
22. Australian Government, Department of Health. Chronic disease: Chronic diseases are leading causes of death., disability in Australia, http://www.health.gov.au/internet/main/publishing.nsf/Content/chronic (last updated: 26 September 2012)
23. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83 (1945)
24. Shapiro, S. S., Wilk, M. B.: An analysis of variance test for normality (complete samples). Biometrika 52(3–4), 591–611 (1965)
25. Razali, N., Wah, Y. B.: Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors., Anderson-Darling tests. Journal of Statistical Modeling., Analytics 2(1), 21–33 (2011)
26. Suominen, H. and Ferraro, G.: Noise in speech-to-text voice: Analysis of errors and feasibility of phonetic similarity for their correction. In Karimi, S. and Verspoor, K. (eds.) Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013), pp. 34–42. Association for Computational Linguistics (ACL), Stroudsburg, Brisbane, QLD, Australia. (2013)
27. Powers, D. M. W. Applications and explanations of Zipf's law. In Powers, D. M. W. (ed.): Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning (NeMLaP3/CoNLL'98), pp. 151–160. ACL, Stroudsburg, PA, USA (1998)
28. Devine, E. G., Gaehde, S. A., Curtis, A. C.: Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. JAMIA 7(5), 462–468 (2000)
29. Al-Aynati, M. M. and Chorneyko, K. A.: Comparison of voice-automated transcription and human transcription in generating pathology reports. Achieves of Pathology and Laboratory Medicine 127(6), 721–725 (2003)
30. 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)
31. Pedersen, T.: Empiricism is not a matter of faith. Computational Linguistics 34(3), 465–470 (2008)
32. Organisation for Economic Development (OECD): OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD, Danvers, MA, USA (2007)
33. Estrin, D. and Sim, I.: Open mHealth architecture: An engine for health care innovation. Science 330(6005), 759–760 (2010)

34. 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)
35. 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. JAMIA 18(5), 540–543 (2011)
36. Huang, C.-C. and Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics May 1 (2015)
37. 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)
38. 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)
39. 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 de-identification of free-text medical records. BMC 8, 32 (2008)
40. 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. JAMIA 20(2), 342–348 (2013)
41. 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)
42. 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, c181and Trials 11, 9 (2010)
43. 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-Chancellors' Committee, Canberra, ACT, Australia (2007 updated 2014)
44. 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 (2012)

Appendix 1: Helpful Tips

Running SCTK

The command

```
bin/sclite -r reference.txt -h reference.txt trn -i spu_id -o all
```

should produce perfect results by using the formatted reference standard both as a reference standard (`-r reference.txt`) and speech-recognized (or hypothesized `-h`) text. The command

```
bin/sclite -r reference.txt -h dragon_nursing.txt -i spu_id -o all
```

uses the formatted reference standard and formatted Dragon output, and hence should produce the aforementioned document-specific and overall evaluation results. The `-i spu_id` option refers to the *transcription* (TRN, i.e., a TXT file where each paragraph captures a handover document, followed by *(documentID_V)* with *_V* specifying the person whose voice/speech is recognized) formatted input files (with the default *extended American Standard Code for Information Interchange* encoding and the default *GNU diff* alignment) to pair the reference standard with the speech-recognized documents. The `-o all` option results in not only the evaluation results as percentages and raw numbers but also more details for analyzing correctly and incorrectly detected text patterns.

Removing punctuation and changing to ASCII

```
#!/bin/bash

for file in *.txt
do
 iconv -f UTF8 -t ASCII//TRANSLIT//IGNORE "$file" > "$filemod.txt"
 tr '[:punct:]|\n|\r' ' ' < "$filemod.txt" > "$file"
 echo " ("$file | cut -d '.' -f 1 | tee -a "$file"
 echo "_V)" | tee -a "$file"
 tr -d '\n|\r' < "$file" > "$filemod.txt"
 cp "$filemod.txt" "$file"
 rm $filemod.txt
done
cat *.txt > output.txt
iconv -f UTF8 -t ASCII//TRANSLIT//IGNORE output.txt > reference.txt
tr 'V)' 'V)\n' < reference.txt > output.txt
cp output.txt reference.txt
rm output.txt
```