Crowdsourced Continuous Improvement of Medical Speech Recognition

W. Salloum, E. Edwards, S. Ghaffarzadegan, D. Suendermann-Oeft, and M. Miller

EMR.AI Inc, 90 New Montgomery St #400,
San Francisco, CA 94105, USA

{wael.salloum, erik.edwards, shabnam.ghaffarzadegan, david, mark.miller}@emr.ai

Abstract

We describe a method for continuously improving the accuracy of a large-scale medical automatic speech recognizer (ASR) using a multi-step cycle involving several groups of workers. The paper will address the unique challenges of the medical domain, and discuss how automatically created and crowdsourced input data is combined to refine the ASR language models. The improvement cycle helped to decrease the original system's word error rate from 34.1% to 10.4%, which approaches the accuracy of human transcribers trained in medical transcription.

Introduction

Dictated medical reports pose many challenges to automatic speech recognition (ASR) due to the linguistic complexity of the domain, the acoustic environment, and the way speakers handle dictation. First, highly complex, domain-specific medical terminology including thousands of drug names render the use of standard language models ineffective. Hence, medical dictation faces a significant out-of-vocabulary challenge (see Table 2 below where a language model trained on more than 100 million tokens results in more than a quarter of singletons in the vocabulary). Second, a multitude of different dictation devices (such as PSTN telephony, Dictaphone, SpeechMike, Digital Voice Recorders (DVRs), to name a few), the background noise in hospitals, hesitations and interruptions, and side conversations affect quality and intelligibility of the recording. Furthermore, the nature of medical report dictation creates a new genre of speech caused by the fact that doctors are not talking to a human and sometimes are not even aware of the fact that transcribers (and/or ASR) are listening to the audio to transcribe it, but assume that the recordings are merely stored for auditing purposes. After a long day of work, doctors often speak hastily, producing fast connected sentences which lack clear juncture, boundaries, or formatting commands. Even itemized lists are often spoken in rapid succession that is unrevealing of logical boundaries. Yet at other times, long pauses are inserted in the middle of sentences and intervals of unfocused speech can be found or prolonged sequences of hesitations.

In this work, we present a system that takes in an audio recording of a medical report and produces a final report formatted to meet the customer's requirements. The system consists of two steps, ASR and automatic formatting. While the ASR system produces a transcription (*ASR hypothesis*) matching the pronounced words in the input audio as accurately as possible, the auto-formatting step transforms the ASR hypothesis into a draft report whose format resembles that of the final letter. Auto-formatting involves post-processing phrases (such as numbers and dates) and introducing sentence boundaries, paragraphs, enumeration and bullet lists, physician normals, sections and headings, etc. Furthermore, preamble statements (e.g. "This is Doctor John Doe dictating") are removed, directives to transcribers (e.g., "thank you for writing this", "I'm sorry, remove this") are skipped, and explanatory phrases (e.g., "first name Kevin K E V for Victor I N") and repeated words (e.g., "she uh she") are handled. Moreover, auto-formatting corrects partial phrases in context (e.g., "temp" to "temperature"); identifies commands and inserts sentences (e.g., "insert my closing statement"); reorders certain sentences; and extracts standard report fields such as date of birth, date of service, patient name, doctor name, and type of visit (e.g., "orthopedic followup evaluation").

To continuously improve the performance of the system on the previous two tasks as we receive more and more data from customers, we implemented a Continuous Improvement Cycle, inspired by similar techniques in large-scale spoke dialog systems (Suendermann et al. 2009; Suendermann and Pieraccini 2013), which engages a crowd in the process of audio transcription and report formatting. The use of medical reports with patient information dictates the use of a *private crowd* where workers sign a non-disclosure agreement protecting the content, and the data can be shared back and forth using HIPAA-compliant channels. Also, the challenges of medical dictation presented above apply to humans as well, and thus the use of a well-trained private crowd is essential to insure the quality of their work and control their throughput to insure the delivery of final medical reports on time.

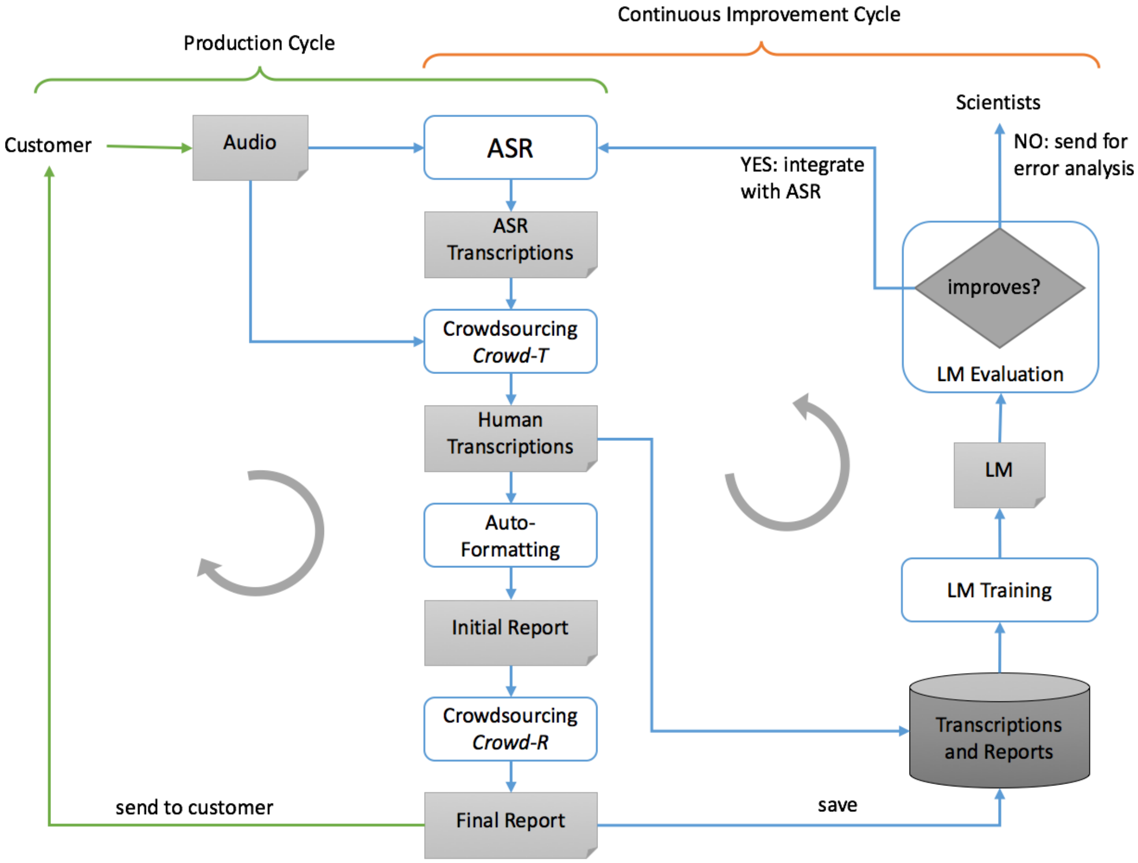

Figure 1: Production Cycle and Continuous Improvement Cycle.

Crowdsourced Continuous Improvement Cycle

Before the introduction of ASR, medical transcription services depended fully on human transcribers. Traditionally, medical transcribers are trained to receive an audio recording of a medical report and produce a full-fledged final report directly from the audio. In contrast, our approach is based on a *crowdsourced transcription process* deploying two distinct groups of transcribers. The first group is referred to as ***Crowd-T*** and corrects the ASR hypotheses producing an accurate one-to-one transcription of the audio with all numerals, hesitations, commands, etc. spelled out. The other group, ***Crowd-R***, generates formatted final reports, working off the output of the auto-formatting module. We integrate these two groups with our system's *Production Cycle* which consists of the following steps (shown on the left side of Figure 1):

1. The customer's client software uploads a dictated report audio file that the ASR system transcribes.
2. Both audio and ASR hypotheses are collected and sent to Crowd-T for human correction. Starting from automatic transcriptions not only saves time and reduces costs, but it also improves the transcription quality. We found that the ASR system often recognizes words that even the trained human transcriber would have a hard time understanding.
3. The human transcriptions are then passed to our auto-formatting module that produces an initial report by 1) post-processing tokens and phrases (e.g., "three point five over five period" to "3.5/5."); 2) removing unneeded sentences and inserting standard statements (e.g., closing statement and headers); 3) restructuring content into titles, sentences and paragraphs; and 4) extracting standard report fields such as date of service and patient name.
4. The initial report is then sent to Crowd-R for quality assurance where a worker corrects formatting errors and produces the final report that is sent back to the customer.

As dictated medical reports keep flowing into our system, it is essential that we continuously adapt our models to new language and keep improving the quality of the various system components. Since the crowd costs are covered by the Production Cycle, we build on top of it by implementing a Continuous Improvement Cycle (shown on the right side of Figure 1) that shares the audio-to-report pipeline, included in and paid for by the Production Cycle, and adds the following steps:

1. The text files produced by Crowd-T and Crowd-R are saved to the transcriptions and reports database. The collection of a certain minimum number of transcriptions

launches the language model (LM) training process which builds an LM from select files in the database.

2. The LM is evaluated on standard development sets and compared to the performance of previous LMs. If the results are equal or better, the new LM is deployed to the production ASR system; if not, partly automatic[1], partly manual error analysis is performed on the new data to identify the root cause of the deterioration.

We treat transcriptions and text reports as two different genres of text. Transcriptions capture the way people talk. They are aligned one-to-one to the audio, where numbers, dates, punctuation marks, formatting instructions (e.g., “new header”, “title”, “paragraph”, “next number”) are spelled out to match the audio. Reports, on the other hand, follow standards typical for out-patient letters. They are formatted and structured in a way that is suitable for reading and fast lookup of information (e.g., fields for Name and Date of Birth). Since ASR systems expect audio as input, the language model must be trained on transcribed text. Final reports, however, are easier to acquire since they are the standard format used by electronic medical record (EMR) systems, while one-to-one transcriptions are either discarded or not generated in the first place. Given the wealth of textual reports we have access to, it seems intuitive to leverage their use in training the LM; however, they need to be transformed to transcription style. This process is the reverse of the auto-formatting process, and thus has similar rules and models. After automatically transcribing select reports from the transcriptions and reports database, we use the output along with select human transcriptions to train our language model.

It is worth mentioning that the two-step crowd transcription process facilitates the creation of training data for three components of our system. The first is the language model being the primary focus of the present paper. The second is the acoustic model which can be built and adapted using the transcriptions along with the input audio. The third is the auto-formatting component and its reverse which can be refined by learning from the relationship between transcriptions and final reports. We reserve the discussion of the latter two components for a separate publication.

Results and Discussion

The ASR system we are using is based on a state-of-the-art stack with 40-dimensional MFCCs, deltas and delta-deltas; fMLLR, ivectors, SAT, GMM-HMM pre-training, and a DNN-based acoustic model trained on hundreds of hours medical dictation audio. The language model uses four-grams with Kneser-Ney smoothing and interpolation to minimize perplexity.

Before launching the Continuous Improvement Cycle, we created a baseline ASR system with 106 million tokens of general medical reports produced by the reverse auto-formatting component. Corpus statistics are shown in Table 2.

System	ASR WER
with LM trained on general medical reports	34.1%
with LM adapted to target population	15.5%
after multiple cycles of tuning	**10.4%**

Table 1: ASR word error rate (WER) results on the test set.

	Gen. med. reports	Target population
# tokens	106M	30M
# types	201K	68K
# singletons	54K	15K
% singletons	27%	23%

Table 2: Corpus statistics.

The test set consists of 19.9 hours of dictations of about 180 physicians speaking US-English. The baseline system, i.e. the system before launching the Continuous Improvement Cycle, produced a word error rate (WER) of 34.1%, see Table 1. This is in line with publications on similar domains, for example medical question answering (Liu et al. 2011), who reported word error rates on spoken clinical language of between 30.5% and 69.1%.

The first major improvement cycle made use of a set of over 30,000 reports of the aforementioned 180 physicians (the “target population”). The resulting system achieved a substantial performance improvement resulting in a WER of 15.5%.

Further enhancing the data, tuning language model smoothing weights to minimize perplexity, tuning the acoustic scale, the ivector extraction window, the feature vector dimensionality, etc. as well as exploring several language model interpolation techniques resulted in further reduction of the WER to 10.4%. This result is close to human performance on the medical transcription task. On a similar database, we have measured a human WER of between 6 and 10%.

Conclusion

The Continuous Improvement Cycle we have described in the present paper is tailored primarily towards language model adaptation and tuning and has proved very effective without creating manual overhead in addition to what the medical transcripion pipeline requires anyway. We have shown that a medical dictation system that was originally trained on over 100 million tokens could be improved from an original 34.1% to 10.4% by rigorously making use of the Continuous Improvement Cycle. Further enhancements include acoustic model adaptation as well as adjustments to the auto-formatting component and its reverse.

Related Work

Although medical-domain ASR has been reported in some form since the 1980s (Leeming et al. 1981; Akers 1986; Matumoto et al. 1987), there is surprisingly little precedent for the work reported here. In fact, all work prior to 1999 used single-word as opposed to continuous speech recog-

[1] An automatic error analysis is performed by evaluating the perplexity of each new file against the previous LM in order to isolate unconventional cases for manual error analysis.

nition. Early works on continuous medical speech recognition (Hundt et al. 1999; Zafar, Overhage, and McDonald 1999; Devine, Gaehde, and Curtis 2000) immediately recognized the importance of including medical domain-specific terminology in the statistical language model. However, the physicians (usually radiologists) were themselves enlisted to provide manual corrections to update the ASR lexicon. This procedure is untenable for anything except small-scale work within a single hospital department. Only gradually in the 21st century have a handful of studies begun to use non-physician transcriptions for language model training, and mostly in the single domain of radiology, e.g. (Paulett and Langlotz 2009), although a small number of single-domain systems have been reported elsewhere (dermatology: (Smith 2002); temporomandibular disorder: (Hippmann et al. 2010)). Our language model methodology scales to larger volumes of data from multiple subspecialties, adapting to each specific domain as well as to speaker and hospital-specific characteristics. Among the few publications on speech recognition on medical corpora is the work by (Cao et al. 2011) and (Liu et al. 2011) on clinical question answering. Comparable NLP motivations are found in the extensive work in medical AI and NLP since the 1970s (reviews: (Clancey and Shortliffe 1984; Cai et al. 2016; Pons et al. 2016)), which work directly from text records (as opposed to voice entry). However, we find a surprising absence of sophisticated AI or NLP methodology in medical ASR, other than the aforementioned studies of Cao, Liu, and colleagues and a small number of radiology studies (Paulett and Langlotz 2009; Ringler, Goss, and Bartholmai 2015). Finally, several tens of papers are written about medical ASR from an administrative, sociological, or economic point of view. These generally conclude in favor of the efficiency and cost-effectiveness of voice vs. manual data entry, but are not covered here (for reviews, see (Johnson et al. 2014; Hammana et al. 2015; Lyons et al. 2016)).

References

Akers, G. 1986. Using Your Voice: Speech Recognition Technology in Medicine and Surgery. *Clin. Plast. Surg.* 13(3).

Cai, T.; Giannopoulos, A.; Yu, S.; Kelil, T.; Ripley, B.; Kumamaru, K.; Rybicki, F.; and Mitsouras, D. 2016. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics* 36(1).

Cao, Y.; Liu, F.; Simpson, P.; Antieau, L.; Bennett, A.; Cimino, J.; Ely, J.; and Yu, H. 2011. AskHERMES: An Online Question Answering System for Complex Clinical Questions. *J. Biomed. Inform.* 44(2).

Clancey, W., and Shortliffe, E. 1984. *Readings in Medical Artificial Intelligence: The First Decade*. Reading, USA: Addison-Wesley.

Devine, E.; Gaehde, S.; and Curtis, A. 2000. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *J. Am. Med. Inform. Assoc.* 7(5).

Hammana, I.; Lepanto, L.; Poder, T.; Bellemare, C.; and Ly, M. 2015. Speech Recognition in the Radiology Department: A Systematic Review. *HIM J.* 44(2).

Hippmann, R.; Dostalova, T.; Zvarova, J.; Nagy, M.; Seydlova, M.; Hanzlicek, P.; Kriz, P.; Smidl, L.; and Trmal, J. 2010. Voice-Supported Electronic Health Record for Temporomandibular Joint Disorders. *Methods Inf. Med.* 49(2).

Hundt, W.; Stark, O.; Scharnberg, B.; Hold, M.; Kohz, P.; Lienemann, A.; Bonel, H.; and Reiser, M. 1999. Speech processing in radiology. *Eur. Radiol.* 9(7).

Johnson, M.; Lapkin, S.; Long, V.; Sanchez, P.; Suominen, H.; Basilakis, J.; and Dawson, L. 2014. A Systematic Review of Speech Recognition Technology in Health Care. *BMC Med. Inform. Decis. Mak.* 14(94).

Leeming, B.; Porter, D.; Jackson, J.; Bleich, H.; and Simon, M. 1981. Computerized Radiologic Reporting with Voice Data-Entry. *Radiology* 138(3).

Liu, F.; Tur, G.; Hakkani-Tur, D.; and Yu, H. 2011. Towards Spoken Clinical-Question Answering: Evaluating and Adapting Automatic Speech-Recognition Systems for Spoken Clinical Questions. *J. Am. Med. Inform. Assoc.* 18(5).

Lyons, J.; Sanders, S.; Cesene, D.; Palmer, C.; Mihalik, V.; and Weigel, T. 2016. Speech Recognition Acceptance by Physicians: A Temporal Replication of a Survey of Expectations and Experiences. *Health Informatics J.* 22(3).

Matumoto, T.; Iinuma, T.; Tateno, Y.; Ikehira, H.; Yamasaki, Y.; Fukuhisa, K.; Tsunemoto, H.; Shishido, F.; Kubo, Y.; and Inamura, K. 1987. Automatic Radiologic Reporting System Using Speech Recognition. *Med. Prog. Technol.* 12(3-4).

Paulett, J., and Langlotz, C. 2009. Improving Language Models for Radiology Speech Recognition. *J. Biomed. Inform.* 42(1).

Pons, E.; Braun, L.; Hunink, M.; and Kors, J. 2016. Natural Language Processing in Radiology: A Systematic Review. *Radiology* 279(2).

Ringler, M.; Goss, B.; and Bartholmai, B. 2015. Syntactic and Semantic Errors in Radiology Reports Associated with Speech Recognition Software. *Health Informatics J.*

Smith, K. 2002. A Discrete Speech Recognition System for Dermatology: 8 Years of Daily Experience in a Medical Dermatology Office. *Semin. Cutan. Med. Surg.* 21(3).

Suendermann, D., and Pieraccini, R. 2013. Crowdsourcing for industrial spoken dialog systems. In Eskenazi, M.; Levow, G.; Meng, H.; Parent, G.; and Suendermann, D., eds., *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Hoboken, USA: Wiley.

Suendermann, D.; Liscombe, J.; Evanini, K.; Dayanidhi, K.; and Pieraccini, R. 2009. From Rule-Based to Statistical Grammars: Continuous Improvement of Large-Scale Spoken Dialog Systems. In *Proc. of the ICASSP*.

Zafar, A.; Overhage, J.; and McDonald, C. 1999. Continuous Speech Recognition for Clinicians. *J. Am. Med. Inform. Assoc.* 6(3).