A Natural Language Correction Model for Continuous Speech Recognition[1]

Tomek Strzalkowski and Ronald Brandow
GE Research & Development
1 Research Circle
Niskayuna, NY 12309
{strzalkowski, brandow}@crd.ge.com

Summary

We have developed a method of improving and controlling the accuracy of automated continu speech recognition through linguistic postprocessing. In this approach, an output from a spe recognition system is passed to a trainable Correction Box module which attempts to locate repair any transcription errors. The Correction Box consists of a text alignment program, a cor tion-rule generator, and a series of rule application and verification steps. In the training phase, correction rules are learned by aligning the recognized speech samples with their original, f correct versions, on sentence by sentence basis. Misaligned sections give rise to candidate c text-free correlation rules, e.g., from ⇒ frontal ; there were made ⇒ the remainder , etc. Validation agains text corpus leads to context-sensitive correction rules, such as from view ⇒ frontal view . The systen applied to medical dictation in the area of clinical radiology.

1. INTRODUCTION

In this paper we describe a method of improving the accuracy of automated speech recognit through text-based linguistic post-processing. The basic assumption of our approach is tha majority of transcription errors can be attributed to either some inherent limitations of the l guage model employed by a speech recognition system, or else to the specific speech patterns c speaker or a group of speakers. Many advanced speech recognition systems use trainable l guage models that can be optimized for a particular speaker (speaker-adaptable, or speaker-in pendent) as well as for a specific sublanguge usage (e.g., radiology). This optimization necessary to achieve a respectable level of recognition accuracy, however, it may not guaran consistently high-accuracy performance due to the limited capabilities of the underlying langu model, usually a 2- or 3-gram HMMs. Our method is to take a reasonably accurate transcripti (perhaps 70-90% word accuracy) and automatically develop a correction filter that would ass

1. This research is based upon work supported in part under a cooperative agreement between the National Insti of Standards and Technology Advanced Technology Program (under the HITECC contract, num 70NANB5H1195) and the Healthcare Open Systems and Trials, Inc. consortium.

consistently the highest possible performance. Unlike other approaches (e.g., Sekine & Grishm 1996) that attempt to choose from among alternative transcriptions based on syntactic and/or l ical well-formedness, our method is to actually identify and correct transcription errors in the S output.

We would like to stress that while the experiments described in this paper are relatively mod and preliminary, the system we designed is robust and fully automatic: there is no human int vention involved.

2. MOTIVATION

In clinical radiology, an on-duty radiologist reviews and reads dozens of images a day: x-ray fi as well as digital imagery such as CTIs and MRIs, etc. A radiology department at a medium si hospital handles hundreds of images every day. In a typical situation, an on-call radiologist exa ines a series of images and dictates his or her findings into a recording device, except in em gency situations where the consultation may be read directly to an attending physician. Voi recordings are then sent for transcription, usually by outsourcing, and return as text reports for t radiologist's signature. The turnaround of this process may take anywhere from a few hours several days, depending upon the priority of the request, and the efficiency of the transcripti service. For a radiologist dictating dozens of reports a day, it is practically impossible to truly v ify the transcription arriving for his signature a few days later - that is verify beyond a surfa "makes-sense" assessment. Therefore, an automated transcription system that can generate printed report within seconds after being dictated, represents a substantial improvement in quali of this process, while eliminating costly manual labor.

The above argument is of course conditioned upon the quality and reliability of automated spee transcription process. It is estimated that a continuous speech recognition system must maintair consistently very high-level accuracy. [The exact figure is subject to verification in clinical te currently underway. To give an example: a 95% accurate speech recognition means on average or 2 errors per a brief 20-40 word report]. Today's automated speech recognition systems, wh quite advanced, are not sufficiently accurate for this application. We believe, however, that a su stantial improvement of automatic transcription quality is possible by augmenting ASR capabi ties with natural language processing techniques. Our initial experiments suggest that this m indeed be possible. It remains to be seen whether the resulting system can be used in a real radi ogy dictation environment.

For the experiments reported in this paper, we used a commercially available continuous spee recognition system, equipped with extensive radiology language model based on phoneme- a word-level HMMs, and specifically designed for dictation. Important selection criteria includ advertised accuracy (95%), availability, compatibility, user interface type, and cost.

3. OUTLINE OF THE APPROACH

The basic premise of our approach is to utilize linguistic and sublanguage-specific informatic even speaker-specific features, in order to improve the accuracy of speech transcription. For t experiments described in this paper, we considered the speech recognition system as a black bc that is, our efforts were directed at correcting the transcription errors in a post-processing mod

rather than to improve the initial transcription. One advantage of this two-pass approach is independence from any particular SRS, and indeed our Correction Box (C-Box) module can used as a back-end of any speech system. Nonetheless, there are other possibilities as well, we mention them briefly in the Future Directions section.

The main goal of the C-Box approach is to generate a collection of context-sensitive text-ba correction rules, in the form of $XLY \Rightarrow XRY$, where X and Y are context word patterns, L is a w pattern in an erroneous transcription string, and R is a replacement string correcting transcript errors in L. In order to generate the correction rules we require both training and validation st to optimize the C-Box performance with respect to a specific sublanguage as well as a speake a group of speakers. Therefore, the C-Box approach consists of the following sub-processes: collecting training data, (2) aligning text samples, (3) generating correction rules, (4) validat correction rules, and (5) applying correction rules.

4. DETAILED ALGORITHM

In this section we discuss the major steps in the C-Box algorithm. Some further details are cussed in subsequent sections.

Step 1. The process is started by collecting a sufficient amount of training text data. The train data must contain parallel text samples: a manually verified true transcription, and the output the ASR system on the same voice input. We estimate that about a day worth of dictation is su cient to initiate the training process, some 300 reports or about 150 KBytes of text. In addition fair-sized text corpus of historical, fully verified transcriptions is required. In the experime described in this paper we used about 1000 reports (400 KBytes) of parallel text, plus an ad tional 5881 correct transcriptions corpus (total 3.25 MBytes). The training was performed on reports, and tested on 200 reports.

Step 2. The parallel text samples are aligned, sentence by sentence on the word level. This is b achieved using content-word islands, then expanding onto the remaining words, while using so common-sense heuristics in deciding the replacement string correspondences: misaligned s tions while different in print, are nonetheless close "phonetic variants". We use character-le alignment for support, especially if the replacement scope is uncertain. Additionally, we allow known spelling variants to align, e.g., "1" for "one", or "Xray" for "X - ray" or "X-ray", etc.

Step 3. When a misalignment is detected, the system lexicon is automatically checked for mi ing words, i.e., we check to see if any given transcription error arose because the speaker had u some word or words that are unknown to the speech recognition system. This is accomplished looking up the misaligned words in a broad-coverage medical terminology lexicon[1]. Since s words will be eventually entered into the SRS lexicon, thus eliminating transcription errors, decided not to generate correction rules for them, at this time.

Step 4. Misaligned sections give rise to preliminary context-free string replacement "rule $L \Rightarrow R$, where L is a section in the ASR output, and R is the corresponding section of the true tr

1. We used UMLS Lexicon and Metathesaurus, available from the National Library of Medicine, as well as a cor mercial Radiology/Oncology spell-checker.

scription. For example, the replacement rules there were made ⇒ the remainder and this or ⇒ the support obtained by aligning the following two sentences:

s1: THERE WERE MADE OF THIS OR LINES AND TUBES ARE UNCHANGED IN POSITION.

s2: THE REMAINDER OF THE SUPPORT LINES AND TUBES ARE UNCHANGED IN POSITION.

Context-free "rules" are derived from perceived differences in alignment of a specific pair of s tences, but they would not necessarily apply elsewhere.

Step 5. The candidate rules derived in the previous step are validated by observing their appli bility across the training collection (which consists of the parallel text training data as well a much larger radiology text corpus). To do so, we collect all occurrences of string L in the A output (e.g., "there were made"), and determine how many times the rule $L \Rightarrow R$ is supported in parallel sample. This gives rise to the support set for a given rule, which we will denote $SP(L \Rightarrow$ The remaining occurrences of L constitute the refute set, $RF(L \Rightarrow R)$.

We also consider substrings of L for validation purposes, e.g.:

s1: AT THE TRACK OF TUMOR ON CHANGE IN ...

s2: ENDOTRACHEAL TUBES ARE UNCHANGED IN ...

can be used to validate the rule on change ⇒ unchanged, even though the misaligned sections are mu longer here. In fact, one way of proceeding is to work with shorter-L rules first. This in turn le to breaking down some unwieldy and unlikely candidate rules, such as the one that may ar from aligning two or more consecutive transcription errors together, as illustrated by the abc example.

The initial validation is based on estimated distribution of L within SP and RF sets, and it m look as follows:

L=FUSION	COUNT=43	MIN WEIGHT=0.84
R=EFFUSION	COUNT=41	
R=EFFUSIONS	COUNT= 1	
R=EFFUSION OR	COUNT= 1	

This says that the word FUSION has been mistakenly generated 43 times in the training sample the SRS output, and this constitutes at least 84% of its total observed occurrences. That is, the were 8 other *correct* occurrences of FUSION within the corpus of all valid reports (i. $\frac{43}{43+8} = 0.84$), and at best all of these would translate correctly when read into the SRS. Out of t 43 cases, 41 are mistranscriptions of the word EFFUSION, therefore the context-free ru fusion ⇒ effusion could be proposed. This rule is not perfect, since it will potentially miscorrect oth occurrences of FUSION, but its application can be expected to reduce an overall transcripti error rate on similar data samples.

Step 6. In many cases, clear cut context-free rules, like the one given above are hard to come b Clearly, rules with validity weights of 50% or less are useless, but even those below 75-80% m

be of little value. One possible way to produce higher quality rules is to specialize them by ad context. In other words, we refine rules by fitting them more closely to the existing evidence. ' can be done by identifying contrasting features within the text surrounding L's occurrences could help us to better differentiate between SP and RF sets. If such a feature or features found, they will be used to restate $L \Rightarrow R$ as a context-sensitive rule, XLY ⇒ XRY, where X an are context features. The context is added, initially words, then possibly some non-terminals, element at a time, on either side of L. The revised rules are re-validated, and the cycle is repe until no further progress is possible.

As an example, consider the following pair of sentences:

s1: PORTABLE **FROM** VIEW OF THE CHEST.

s2: PORTABLE **FRONTAL** VIEW OF THE CHEST.

The misalignment gives rise to the context-free correction rule from ⇒ frontal. As we validate rule against the training corpus we find some supporting evidence, but also many cases where rule can't apply, like the one below:

... ARE UNCHANGED IN POSITION **FROM** THE PRIOR EXAMINATION.

However, adding a one-word context of the word VIEW, i.e., replacing the context-free rule ' a context-sensitive from view ⇒ frontal view produces a very good correction rule. More advar possibilities include adding non-terminals to the rules, as in had CD hours ⇒ at CD hours, where stands for any number (here the word tagged with CD part-of-speech).

Step 7. The above process may lead to alignment re-adjustments, as suggested in step 5. U re-alignment, additional rules may be postulated, while other rules may be invalidated. ' would require another cycle of rule validation. This, again, is repeated until no further progre possible, that is no further changes to the rule set result. The final set of context-sensitive cor tion rules is generated.

5. EXPERIMENTAL RESULTS

The experimental data was obtained from the University of Maryland Medical Center in B more, which is one of the clinical sites used in this project. The validated transcriptions have b extracted from the hospital database by the hospital personnel, and then sanitized to remove patient information such as names, addresses, etc. At the time this report is written, we colle nearly 7000 transcribed dictations, all in the area of chest X-ray. Chest X-ray is the most preva form of radiology, and we decided to start with this sub-area because of its the largest pote practical significance.

The sanitized reports were subsequently re-dictated through the automated speech recogni system in order to obtain parallel samples of automated transcription. The redictation was d over a period of several months by a final-year radiology resident at Albany Medical Cente native North American English speaker. At the time this paper is prepared, some 1000 rep have been redictated. Clinical tests of the system equipped with the C-Box that are startin early 1997 will provide additional speakers.

Generally, we observed significant word error rates in automated speech recognition, in so cases as high as 38%, with the average of 14.3%. This is substantially higher than the advertis 5% error rate. Before starting re-dictation, the speaker underwent a few hours training sessi learning how to use the system, and having his voice patterns incorporated into the langua model (the system we use is speaker adaptable). The above numbers therefore represent an op mal performance of the system for this speaker, although there are some hard-to-measure mitig ing considerations. For example, the radiology reports used in these experiments were read by AMC resident, who while obviously familiar with the subject matter, also pointed out some fi vocabulary and style differences between AMC and UMMC Baltimore, where the reports we produced. This could potentially have an impact at SRS performance. It should be noted tha typical chest X-ray dictation report is quite short, from a few lines to a few paragraphs, and is d tated quite rapidly in anywhere from 15 seconds to a few minutes.

Preliminary experiments with context-free rules have already shown interesting results: noticed that the average word error rate decreased from 14.3% to 11.3% (a 21% reduction) o test sample after running it through a C-Box equipped with only a few CF rules. This C-Box w trained on 800 reports (0.3 MByte) and tested on 200 reports (92 KBytes).

Below is a sample radiology report, its automated transcription version, and the effect of a part correction. Note that only context-free rules are used; a context-sensitive correcti indication colon ⇒ indication : would fix the problem in the first line.

Original ASR Transcription: (errors highlighted)

indication **colon and trachea to place.**

the endotracheal tube is in **size factor** position. there **is and** re-expansion of the right upper lobe. mild changes of **the 8th** rds persist bilaterally.

True Transcription:

INDICATION: ENDOTRACHEAL TUBE PLACEMENT.

THE ENDOTRACHEAL TUBE IS IN SATISFACTORY POSITION. THERE HAS BEEN RE-EXPANSION OF THE RIGHT UPPER LOBE. MILD CHANGES OF LATE RDS PERSIST BILATERALLY.

Partially Corrected Transcription: (corrections bold, uncorrected errors italics)

indication *colon* **ENDOTRACHEAL TUBE** place .

the endotracheal tube is in **SATISFACTORY** position. there **HAS BEEN** re-expansion of the right upper lobe. mild changes of *the 8th* rds persist bilaterally.

Correction rules used: and trachea to ⇒ endotracheal tube , size factor ⇒ satisfactory , is and ⇒ has been .

6. OTHER APPROACHES

Natural language processing techniques have been used before to assist automated speech re nition either in constructing better language models, or as post-processors. For example, par speech information has been used to reduce the overall perplexity of the language model. M advanced linguistic methods include re-ranking N-best sentence hypotheses using syntactic lexical well-formedness. A good example is an ongoing effort by Grishman and Sekine (1 1996) at using syntactic parser to pinpoint the correct transcription among the N-best alternat returned by an SRS. Despite the intuitive appeal of their approach, in which transcribed sente are re-ranked using the likelihood or degree of syntactic correctness, they have thus far b unable to obtain noticeable reduction of word error rates, partly due to, as they point out, a lim possible range of improvement: one can only improve N-best ranking if there is a better transc tion among them. Other work with NLP-motivated re-ranking of N-best alternative transcrip include Rayner et al. (1994), Ostendorf et al. (1991), Norton et al. (1992), among others, and also (Hirschman, 1994) and (Moore, 1994) for summary overview. Some alternative possibil of taking advantage of linguistic information in speech recognition are described by e.g., Ku (1992), Murveit and Moore (1990), Maltese and Mancini (1992), and Schwartz et al. (1994).

The post-correction approach has been considered by Ringger and Allen (1996), for a diffe domain (trains scheduling dialog), and using a probabilistic modeling technique rather than rection rules. They report some interesting preliminary results, however these are not dire comparable to ours for two reasons. First, the baseline SRS system used in these experimen much weaker (only about 58% accurate). Second, the reported improvement cover correction vocabulary deficiencies, not only true transcription mistakes due to weaknesses of the langu model.

It should also be noted here that while the above efforts usually attack the more general prob of speech understanding accuracy, in an ad-hoc speech production situation (though limite certain domains such as broadcast news), our solution has been specifically tailored to clin dictation, and wouldn't necessarily apply elsewhere. Some of the application characteristics we take advantage of is relatively limited vocabulary (hence smaller training samples), a lim number of speakers (hence speaker independence less critical), and relatively low perplexit radiology sublanguage (about 20 vs. about 250 for general English as measured for trigram m els (Roukos, 1995)). The C-Box approach does not preclude N-Best techniques - in fact we c sider this as a natural extension of the present method - one of the obvious limitations of present approach is the need for parallel training texts, which may be replaced by multiple al natives.

Overall, the C-Box approach is partly related to error-driven learning techniques as used for p of-speech tagging (Brill, 1992, 1995), and spelling correction (e.g., Golding & Schabes, 1996

Text alignment methods have been discussed primarily in context of machine translation, (e Brown et. al, 1991; Church, 1993; Chen, 1993) and we draw on these here. Rule validatio based in part on the N-gram weighting method described in (Strzalkowski & Brandow, 1996).

7. CONCLUSIONS, LIMITATIONS, AND FUTURE DIRECTIONS

While we are generally pleased with the initial progress of this work, it is still quite early to dr any definite conclusions whether the C-Box method will prove sufficiently robust and effective practical applications. For one thing, we have thus far avoided the question of speaker depe dence. As described, the C-Box method is clearly speaker dependent, that is, the correction rul need to be learned for each new speaker. It remains to be seen if the present solution is acceptab and if a degree of speaker independence can be introduced through rule generalization. Furth research and evaluations are required with context-sensitive correction rules and various sizes the training data. At this time it is also an open question how much improvement can be achiev using this method, i.e., if there is an upper bound, and if so what that is. On the face of it, th seems to depend only on how good rules we can get. In practice, we face limits of rule learnabili due to sparse data, as well as rule interference (i.e., one rule may undo another, etc.). We plan study these issues in the near future.

There are also other possibilities. Some SRS produce ranked lists of alternative transcriptions (1 Best), which can be used to further improve the chances of making only the right correctio Using N-Best sentence hypotheses may also alleviate the need for parallel correct transcriptio in the training sample, as multiple hypotheses can be aligned in order to postulate correction rul Using multiple SRS in parallel may also increase the likelihood of locating and correcting spu ous transcription errors. Finally, we may consider an open-box solution where the informati encoded in C-Box rules is fed back into the SRS language model to improve its baseline perfc mance.

At this time, we made no attempt to address any of the problems related to spontaneous speec such as disfluencies and self-repairs (e.g., Oviatt, 1994; Heeman & Allen, 1994). In dictatio where the speaker normally has an option of backing up and re-recording, such things are less an issue than the word-for-word accuracy of the final transcription, since there are serious liabili considerations to be reckoned with.

Acknowledgements. This research is based upon work supported in part under a cooperati agreement between the National Institute of Standards and Technology Advanced Technolo Program (under the HITECC contract, number 70NANB5H1195) and the Healthcare Open Sy tems and Trials, Inc. consortium. The authors would like to thank all members of the HITEC IMS project at GE CR&D, GEMS, SMS, SCRA, UMMC, Advanced Radiology, and CAMC f their invaluable help, particularly Glenn Fields, Skip Crane, Steve Fritz and Scott Cheney.

REFERENCES

Brill, E. 1992. A simple rule-based part of speech tagger", Proceedings of 3rd Conference Applies Natural Language Processing, Trento, Italy.

Brill, E. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processin A Case study in Part-of-Speech Tagging", Computational Linguistics, vol. 21, No. 4, pp. 543-56

Brown, P., J. Lai, and R. Mercer. 1991. "Aligning sentences in parallel corpora", Proceedings the 29th Annual Meeting of the ACL, Berkeley, pp. 169-176.

Church, K. 1993. "Char_align: A Program for Aligning Parallel Texts at the Character Lev Proceedings of the 31st Annual Meeting of the ACL, Columbus, pp. 1-8.

Chen, S. 1993. "Aligning Sentences in Bilingual Corpora Using Lexical Information", Proce ings of the 31st Annual Meeting of the ACL, Columbus, pp. 9-16.

Golding, A. and Y. Schabes. 1996. "combining Trigram-Based and Feature-Based Methods Context-Sensitive Spelling Correction", Proceedings of 34th Annual Meeting of the ACL, Sa Cruz, pp. 71-78.

Heeman, P. and J. Allen. 1994. "Tagging Speech Repairs", ARPA Human Language Technol Workshop, pp. 187-192.

Hirschman, L. 1994. "The Roles of Language Processing in a Spoken Language Interface" Voice Communication between Humans and Machines, 1994, National Academy of Sciences, 217-237.

Jelinek, F., Merialdo, S. Roukos, and D. Strass. 1991. "A Dynamic Language Model for Spe Recognition", Proceedings DARPA Speech and Natural Language Workshop, pp 293-295.

Kupiec, J. 1992. "Probabilistic Models of Short and Long Distance Word Dependencies in R ning Text", Proceedings DARPA Speech and Natural Language Workshop, pp 290-295.

Maltese and Mancini. 1992. "An Automatic Technique to include Grammatical and Morphol cal Information in a Trigram-based Statistical Language Model", IEEE International Confere on Acoustics, Speech and Signal Processing, pp 157-160.

Moore, R.C. 1994. "Integration of Speech with Natural Language Processing", in Voice Comm nication between Humans and Machines, National Academy of Sciences, pp 254-271.

Murveit, H. and R. Moore. 1990. "Integrating Natural Language Constraints into HMM-ba Speech Recognition", IEEE, pp 573-576.

Norton, L, D. Dahl, and M. Linebarger. 1992. "Recent Improvements and Benchmark Results Paramax ATIS System", Proceedings of DARPA Workshop on Speech and Natural Language.

Rayner, M., D. Carter, V. Digalakis, P. Price. 1994. "Combining Knowledge Sources to Reo N-Best Speech Hypothesis List", Proceedings DARPA Speech and Natural Language Worksh pp. 217-221.

Ostendorf, M. et. al. 1991. "Integration of Diverse Recognition Methodologies Through Reev ation of N-best Sentence Hyphotheses", Proceedings of DARPA Speech and Natural Langu Workshop.

Oviatt, S. 1994. "Predicting and Managing Spoken Disfluencies During Human-Computer In action", ARPA Human Language Technology Workshop, pp. 222-227.

Ringger, E. K, and J. F. Allen. 1996. "Error Correction via a Post-Processor for Continuc Speech Recognition." In Proc. of ICASSP-96, IEEE-96.

Roukos, S. 1995. "Language Representation", In Survey of the State of the Art in Human La guage Technology. Sponsored by NSF, EC, OGI. pp. 35-41.

Schwartz, R., L. Nguyen, F. Kubala, G. Chou, G. Zavaliagkos, and J. Makhoul. 1994. "On Usi Written Language Training Data for Spoken Language Modelling." Proceeedings of Human La guage Technology Workshop. Morhan Kaufmann Publishers, Inc. pp. 94-98.

Sekine, S. and R. Grishman. 1996. "NYU Language Modeling Experiments for the 1995 CS Evaluation", Spoken Language Systems Technology Workshop; New York, NY.

Sekine, S., J. Sterling and R. Grishman. 1994. "NYU/BBN 1994 CSR Evaluation", Spoken La guage Systems Technology Workshop; Austin TX.

Strzalkowski, T. and R. Brandow. 1996. "Spotting Technical Concepts in Natural Langua Text." Proceedings of FLAIRS-96 workshop on Real Natural Language Processing, Orlando, F