Journal of Digital Imaging

Improving the Utility of Speech Recognition Through Error Detection

Kimberly Voll, Ph. D.,[1] Stella Atkins, Ph. D.,[1] and Bruce Forster, M. D.,[2]

Despite the potential to dominate radiology reporting, current speech recognition technology is thus far a weak and inconsistent alternative to traditional human transcription. This is attributable to poor accuracy rates, in spite of vendor claims, and the wasted resources that go into correcting erroneous reports. A solution to this problem is post-speech-recognition error detection that will assist the radiologist in proofreading more efficiently. In this paper, we present a statistical method for error detection that can be applied after transcription. The results are encouraging, showing an error detection rate as high as 96% in some cases.

KEY WORDS: Speech recognition, error detection, radiology reporting, co-occurrence relations, statistical natural language processing, computer-assisted proofreading

INTRODUCTION

The recent improvements of speech recognition (SR) technology have motivated the introduction of automated transcription software in lieu of human transcription. Speech recognition can offer improved patient care and resource management in the form of reduced report turnaround times, reduced staffing needs, and the efficient completion and distribution of reports.[1–3] As the technology comes of age, however, with vendors claiming accuracy rates as high as 99%, the potential advantages of SR over traditional dictation methods are not being realized, leaving many radiologists frustrated with the technology.[4–6]

The primary reason behind this apparent failure is accuracy. A 99%-accurate speech recognizer still averages one error out of every hundred words, with no guarantees as to the seriousness of such errors. Furthermore, actual accuracy rates in the reading room often fall short of 99%. Radiologists are instead forced to maintain their transcriptionists as correctionists, or to double as copy editors, painstakingly correcting each case, often for nonsensical or inconspicuous errors. Not only is this frustrating, but it is a poor use of time and resources. To compound matters, problems integrating with the radiology suite and the introduction of delays have further soured many radiologists on the technology. Those choosing to modernize their reading rooms with SR software are often plagued with difficulties, whereas those continuing to use traditional reporting methods have mixed incentives with respect to upgrading. Nonetheless, the potential benefits to radiology reporting from a hospital administration standpoint continue to motivate the adoption of SR technology. Thus, improving SR dictation is of particular importance.

One potential solution is to improve accuracy through automated error detection. Current research in post-SR error detection has been applied mostly to conversational systems. Statistical methods such as word co-occurrences[7–9] are popular

[1]From the School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.

[2]From the Department of Radiology, University of British Columbia, 2329 West Mall, Vancouver, BC, V6T 1Z4, Canada.

Correspondence to: Kimberly Voll, Ph. D., School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada; Tel: +604-291-4277; Fax: +604-291-3045 e-mail: kvoll@cs.sfu.ca

Online publication 27 March 2007
doi: 10.1007/s10278-007-9034-7

because SR errors "are found to occur in regular patterns rather than at random."[10] Sarma and Palmer[11] use co-occurrence statistics to analyze the context of words in a dialogue query to identify and correct errors. In a similar vein, Inkpen and Désilets[12] use pointwise mutual interpretation, a statistical measure of the independence of two terms, to determine errors in meeting transcripts. To our knowledge, however, no statistical error detection techniques have been applied to post-SR radiology report detection.

Thus, we explore the following hypothesis: As a postprocessing stage, methods in statistical natural language processing can effectively detect recognition errors in radiology reports dictated via automatic speech recognition. As initial support for this hypothesis, our experiment examines one such statistical-based method of error detection: Exploiting the highly regular language used in radiology, we have developed a preliminary system for computer-assisted report editing. By automatically flagging potential errors, the system removes the need for an in-depth reread of the dictated report and reduces the time spent proofreading. If we consider computer-aided diagnosis (CAD) in imaging as a "second pair of eyes" for the radiologist, then by analogy we are proposing a CAD system for error detection in SR-generated reports. This in turn restores much of the original benefit of SR technology in time and money saved. Finally, by creating a postprocessing solution, we can escape the need to deal with proprietary software and recognize the varying needs of reading rooms supporting different vendor software packages.

MATERIALS AND METHODS

Overview

Underlying our approach is the key notion that, by identifying patterns common to error-free reports, we can automatically detect inaccuracies within novel reports. Our pattern analysis has been done via co-occurrence relations,[8,11,13] a statistical method for determining the number of times a word occurs in a specific context. Here, "context" refers to the words surrounding the target word. Given a sufficiently representative training corpus, we can associate words with particular contexts based on that corpus. We can then apply these word–context statistics to determine the probability of a word occurring in a given context in a report. If that probability falls below a certain threshold, the word will be flagged as a possible error, assisting the radiologist during proofreading.

Materials

With permission from the Simon Fraser University Ethics Committee, we have compiled the co-occurrence statistics for 2,700, anonymized magnetic resonance imaging reports, obtained and corrected by the Canada Diagnostic Centre (CDC) in Vancouver, British Columbia, using the Dragon NaturallySpeaking speech-recognition system, version 7.3. This training corpus is split into several training sets: the full 2,700 reports, as well as those obtained from dividing by section and dividing by report type (ie, anatomic region being studied). These divisions reflect the observations that the type of words found in the "Findings" and "Impressions" sections may differ from the "History" section, whereas the type of words found within a knee report, for instance, is not as likely to occur in a report of the shoulder. Thus, by training and testing these separately, there is no risk of dilution from other report types, increasing the accuracy. Note that pathology-based division of reports was not considered at this stage, although, like anatomy-based reports, similar behavior with respect to types of words is expected.

Our final training sets include all reports, reports separated into the "Findings" and "Impressions" sections, and reports of the spine. To ensure adequate statistical representation, we restrict training sets to those containing 800 or more reports. Of the 2,700 reports divided by anatomic region, only "spine" had enough cases to meet the 800-minimum requirement. We generate separate co-occurrence statistics for each training set based on the current context window size.

Stopwords are words with little intrinsic meaning, such as "at" and "the." Typically, these words are found with such high frequency that they lose any usefulness as search terms. In co-occurrence analysis, stopwords are usually omitted because their overabundance in a text can affect the resulting probabilities disproportionately. We observe this convention in the following experiment.

A sample selection of the co-occurrence relations for the word "quadriceps" from the training corpus is provided in Table 1. For example, "quadriceps" occurred in the training corpus 123 times and co-occurred with the term "patellar" 32 times, for a frequency of 32/123=0.26.

Methods

In the testing phase, we compiled a corpus of 20 uncorrected/corrected, anonymized report pairs, also obtained from the CDC using Dragon NaturallySpeaking. For each uncorrected report, we determine the context of each word, calculate the co-occurrences, and apply the appropriate collection of co-occurrence statistics from the training data.

For example, consider the following misrecognized sentence fragment:

> ...possible spondylolysis eye laterally of L5...

We can generate the following co-occurrences for the target word, "eye", with a context window of 2 (up to two words to either side of "eye"):

eye possible
eye spondylolysis
eye laterally
eye L5

Using Bayes' theorem (Eq. 1), we can combine the probability of each word occurring within the context window of the target word, and the probability of the target word itself, where T is the target word and C is the context words. Bayes' theorem is a formula that allows us to calculate conditional probabilities: the probability of an event, A, given the knowledge that another event, B, has already taken place. In simpler terms, this means that the probability of our "event", the target word T, can be calculated in terms of the probability of another "event", the context C. Because the target word and the context are closely related, this is an informative calculation.

Table 1. Co-occurrence Statistics for "Quadriceps"

Term	Context Word	Count	Term Count	Freq.
Quadriceps	Included	1	123	0.01
Quadriceps	Mechanism	1	123	0.01
Quadriceps	Patellar	32	123	0.26
Quadriceps	Tendon	38	123	0.31
Quadriceps	Tendons	50	123	0.41
Quadriceps	Vastus	1	123	0.01

$$P(T|C) = \frac{P(T)P(C|T)}{P(C)}. \qquad (1)$$

The expression $P(T \mid C)$ is read "the probability of T given C." The probability of the target word, $P(T)$, is equal to the probability of occurrence in the training corpus. Because we have already observed the context of the target word, we know that its probability of occurring is constant; thus, we set $P(C)$=1. Finally, we can calculate $P(C \mid T)$, the probability of the context C occurring given the target word T, where C_1,...,C_n represent the context words within C, as follows:

$$\begin{aligned} P(C|T) &= P(T)P(C_1, \ldots, C_n|T) \\ &= P(T)P(C_1|T) \times \ldots \times P(C_n|T) \;\; \text{By joint probability} \\ &= P(T)\prod_{i=1}^{n} P(C_i|T) \end{aligned} \qquad (2)$$

Given that $P(C_i \mid T)$ is equivalent to the probability of the co-occurrence (C_i, T) in the training data, we can now calculate our desired probability, $P(T \mid C)$.

For example, applying Bayes' theorem to the sentence fragment above yields the following:

$$P(\text{eye}|\text{possible, spondylolysis, laterally, L5}) = P(\text{eye}) \times P(\text{possible, spondylolsis, laterally, L5}|\text{eye})$$

We can apply Eq. 2 to calculate P(possible, spondylolysis, laterally, L5 | eye):

$$P(\text{eye}|\text{possible, spondylolysis, laterally, L5}) = P(\text{eye}) \times P(\text{eye}|\text{spondylolsis}) \times P(\text{eye}|\text{laterally}) \times P(\text{eye}|\text{L5})$$

Once we have obtained the value of $P(T \mid C)$ via Bayes' theorem, we can compare it to a threshold value, K, flagging those target words, T, where $P(T \mid C) < K$. Thus, we capture those words in a report whose occurrence in their context window is highly improbable. This improbability reflects the likelihood of a recognition error. To find the actual errors in our test reports, we align the corrected and uncorrected reports, determine any differences, and tag those differences as errors. We then compare these to the flagged errors from our program output to obtain our results: a match is considered a correct detection, or true positive; a flagged error that does not correspond to an actual error is considered a false positive; an error not flagged is considered a false negative.

For example, after applying Bayes' theorem and Eq. 2 to our sample sentence fragment, we have *P*(eye|possible,...,L5)=4.37067E-07, a correspondingly low value that reflects the unlikelihood of "eye" occurring in that context. Assuming an appropriate threshold *K*, this is flagged as an error.

As a final note on the use of Bayes' theorem, as described above, the probability of a target word is defined with respect to the individual probabilities of the words in the surrounding context along with the probability of the word occurring on its own. It is not simply the occurrence of *T* within the scope of some context word, *C*, but a combined measure of the probability of *T* ever occurring (as defined by our training corpus) within range of *C*, along with the probability of *T* occurring overall (again, as defined by our training corpus). In the base case, if *T* has never occurred next to any of the context words, then the overall probability will be zero, whereas similarly, if *T* is a novel word, the overall probability will also be zero. However, the more complex case, where *T* has occurred before and in the environment of one or more of the current context words, is a balance of the frequency of those occurrences with respect to each context word, *C*. If the threshold is set to zero, this simply catches the simple case as described above, and is insufficiently discriminating (although it does offer sufficient results as proof of concept). A carefully chosen threshold value, however, can increase the discriminating power of the system, and thus improve overall accuracy.

All tools were designed in Perl and run on a Mac G4, 1.5 GHz, OS X 10.3.9. All calculations were performed on a context window of size 3, with a threshold value of 0.0001.

RESULTS

Table 2. Postprocessing Outcome of Error Detection on Speech-Recognized Reports

	Accuracy		Corpus Size	
Report Type	Recall (%)	Precision (%)	Training	Test
All	83	26	2,751	20
Findings only	88	31	2,751	20
Impressions only	96	15	2,751	20
"Spine" only	77	35	891	10

The results obtained are shown in Table 2. Recall is a measure of the number of errors correctly detected over the total number of errors actually present; precision is a measure of the number of errors correctly detected over the total number detected. In keeping with our CAD analogy, we can refer to these two terms as the sensitivity and specificity, respectively. Corpus size:training is the number of reports in the training set, whereas corpus size:test is the number of test cases on which the system was run.

Out of 20 test reports, there is an average of 11.9 errors per report, with an average report length of 80.8 words. This represents an average word-error rate (WER) of 15%.

The system is able to identify error candidates in under a minute in all cases, underscoring its viability for real-time use. There is a one-time overhead cost associated with generating the co-occurrence statistics for the training sets. Once generated, however, the database is simply stored and referenced. Regeneration then only occurs if new training data are added. The cost and complexity of referencing the database is dependent upon the number of entries (word–context pairs). As this is simply a lookup task, however, many efficient techniques exist for doing so while ensuring that the system remains viable.

DISCUSSION

The initial results in Table 2 are promising. The recall reflects a moderate sensitivity to errors and a moderately low rate of false negatives. This is especially important, as errors missed could have serious ramifications. In contrast, the precision is low, indicating a high rate of false positives.

There is an interesting discussion to be had regarding the actual utility of the results. In particular, we can ask how useful a recall result of less than 100% is. In keeping with the CAD analogy, the system is a second set of eyes detecting errors that the radiologist may have missed. Arguably, the system proposed here in its current, preliminary stage is suitable only as an assistive device to support the proofreading of the radiologist. The radiologist must remain alert to other errors that may have escaped the error-detection system. As the recall performance increases, the error-detection system becomes one of increasingly accurate report verification. The comfort level with respect to the level of autonomy afforded to such an error-detection system will vary according to many factors, including the radiologist's own beliefs and comfort levels, and the accuracy rate of the system. The higher the accuracy is the lesser the reliance on radiologist proofreading and the increased utility of the system.

In addition, it cannot be forgotten that the recall score is always tempered by the precision score. Obviously, a system that tags every word as an error will enjoy a recall performance of 100%, but zero utility due to a 0% precision score (functionally equivalent to tagging nothing as an error). It can be argued that a lower precision score (high false positives) is less urgent than a lower recall score. While still important overall, false positives remain identifiable by the radiologist and do not affect report quality.

In this study, most false positives were generated by word–context pairs that were not previously encountered in the training data. Thus, we have $P(C \mid T)=0$, which results in $P(T \mid C)=0$ by Eq. 1. Evidence for this is seen in the "Impressions" data set, which typically held the smallest amount of text, and the smallest training set. Correspondingly, it has the lowest precision rate. By increasing the number of reports in the training corpus, however, we can ensure a greater coverage of the terms that typically occur in a radiology report. This will cause the rate of false positives to drop and improve the precision. Although the ideal training corpus would contain every possible context of every possible word in a radiology report, radiology nonetheless does not exhibit a wide variation within reports. A fairly accurate depiction of the possible patterns within a report is feasible with a large enough training set. Interestingly, however, some false positives may be advantageous, indicating rare occurrences that merit closer inspection by the radiologist to ensure there are no mistakes.

Separating the training and testing data by section has a positive impact, although further testing is needed. This result is encouraging, as the "Impressions" section is the section most likely to be read by the referring physician. As mentioned above, the lower precision for "Impressions" is explained by the typically small amount of text in this section. Thus, while separating by type improved recall, overall, the training set was still too small for as effective an analysis and must be followed up with more data.

The rate of error detection, or filtering, is affected by the threshold value, K. Higher values of K mean less filtering and a higher WER, while lower values of K mean greater filtering and a lower WER. In this way, it is possible to increase the recall level to near 100% in exchange for a corresponding loss of precision. Nonetheless, this does allow for some flexibility in balancing between the recall and precision measurements.

The choice of threshold is presently one of trial-and-error experimentation. In extending the CAD analogy, however, we can see that the individual word statistics and their error/nonerror status lend themselves rather nicely to plotting via receiver-operating characteristic (ROC) curves. Doing so would permit a visual analysis of a more complete range of threshold values and allow for further experimentation with respect to other experiment variables such as corpus size and window size. At this stage, ROC analysis remains future work.

An important aspect of this analysis is the omission of stopwords, or low-information-bearing words. These words are ignored because it is often observed that a misrecognized stopword rarely entails a shift in the intended semantics. Exceptions exist, however, such as a substitution of "and" for "at the", that may have more serious consequences in medicine and may prove difficult for human editors to detect. As a result, a more detailed analysis of stopwords is currently underway.

Closely related to the problem of stopwords is the loss of short words with high (relative) semantic load. For example, the omission of the word "no" can have serious ramifications for the meaning of the final report. In many cases, these types of words fall under the stopword category "high frequency" that dilutes the efficacy of statistical-based analysis. In partial answer to this problem, a technique for a hybrid approach to error detection employing multiple methods (statistical and nonstatistical) has been developed that marries the strengths of each method for maximal recall and precision.[14]

Further extensions for this system may include integration with a “talking template”, as proposed by Sistrom.[15] The talking template provides report navigation feedback as audible cues, reducing the so-called look away problem in which radiologists continually look away from the images to visually consult the report being generated. The introduction of speech-recognition systems (SRS) is such that “most radiologists learn to use SRS in a counterintuitive way whereby they interact intensely with the graphical interface to produce reports singly rather than in groups in batches”[16] (p 178). Sistrom suggests that the goal is to defer any proofreading or editing of the report so that multiple reports can be completed at one time in a more efficient “batch mode.” As he observes, “there seems to be fascination bordering on obsession with checking to see if the recognition of the last couple of sentences was accurate”[16] (p 179). With this in mind, the on-the-fly error detection may ultimately provide its feedback as auditory indications that will allow the radiologist to quickly and efficiently deal with such errors (such as redictating). Alternatively, the awareness that the error-detection system is in place may provide radiologists with greater peace of mind and help break the “look away” cycle.

Additional work supporting our hypothesis is also ongoing, including experiments with innovative, nonstochastic methods that rely on syntactic as well as semantic analysis of the text itself, such as conceptual similarity.[14–17] Such techniques will make it possible to evolve beyond just detection to the much harder problem of automated correction. This can include semiautomated correction, whereby the radiologist is presented with intelligent suggestions for correction of recognition errors. Still, as a first attempt at statistical error detection in radiology reports, these results are encouraging and demonstrate the feasibility of postprocessing error detection as a means to recover from the low accuracy of SR. To the authors’ knowledge, error detection in medical SR is a new research area. Therefore, these initial results will provide a starting point for future comparisons and hopefully inspire further work.

Finally, this technique could easily be extended to other areas of medicine that share the same properties of restricted vocabulary seen in radiology, provided an adequate training corpus is available.

CONCLUSIONS

Despite the trend towards automation in the reading room, SR remains a weak alternative to traditional transcription. This is attributable to poor accuracy rates and the wasted resources spent on proofreading erroneous reports. As a partial solution to this problem, we have proposed a post-speech-recognition, statistical error-detection system for radiology. A previously unexplored area of research, this technique shows promise as an effective means to recover from the unacceptable accuracy rates of SR. By flagging potential errors, we can enhance the proofreading process, restoring the benefits of SR in resources saved. The result is a more efficient reading room and an improved experience with SR.

ACKNOWLEDGMENTS

We wish to thank the CDC for their continued support of this research endeavor, as well as Simon Fraser University.

REFERENCES

1. Horii SC, Redfern R, Kundel H, et al: PACS technologies and reliability: are we making things better or worse? Proc SPIE 4685:16–24, 2002

2. Mehta A, Dreyer K, Schweitzer A, et al: Voice recognition—an emerging necessity within radiology: Experiences of the Massachusetts general hospital. J Digit Imaging 11:20–23, 1998

3. Al-Aynati NM, Chorneyko KA: Comparison of voice-automated transcription and human transcription in generating pathology reports. Arch Pathol Lab Med 127(6):721–725, 2003

4. Ranaa DS, Hurst G, Shepstone L, et al: Voice recognition for radiology reports: is it good enough? Clin Radiol 60(11):1205–1212, 2005

5. Marion J: Radiologists’ attitudes can make or break speech recognition. Diagn Imaging Online http://www.superiorconsultant.com/Pressroom/Articles/Diagnostic%20Imaging%20Feb%202002.doc

6. Gale B, Safriel Y, Lukban A: Radiology report production times: voice recognition vs. transcription. Radiol Manage 23:18–22, 2001

7. Jeong M, Kim B, Lee G: Using higher-level linguistic knowledge for speech recognition error correction in a spoken Q/A dialog. In: Proceedings of the HLT-NAACL special workshop on Higher-Level Linguistic Information for Speech Processing, Boston, USA, 2004, pp 48–55

8. Jurafsky D, Martin J: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Englewood Cliffs: Prentice-Hall, 2000

9. Allen JF, Miller BW, Ringger EK, et al: A robust system for natural spoken dialogue. In: Proceedings of the 34th Annual Meeting of the ACL, Santa Cruz, USA, 1996, pp 62–70

10. Kaki S, Sumita E, Iida H: A method for correcting errors in speech recognition using the statistical features of character co-occurrence. In: ACL-COLING, Montreal, Canada, 1998, pp 653–657

11. Sarma A, Palmer D: Context-based speech recognition error detection and correction. In: Proceedings of the HLT-NAACL, Boston, USA, 2004, pp 85–88

12. Inkpen D, Désilets A: Semantic similarity for detecting recognition errors in automatic speech transcripts. In: Proceedings of EMNLP. Association for Computational Linguistics, Vancouver, Canada, 2005, pp 49–56, http://www.aclweb.org/anthology/H/H05/H05-1007

13. Manning CD, Schütze H: Foundations of Statistical Natural Language Processing. Cambridge: MIT Press, 2002

14. Voll K: A Methodology of Error Detection: Improving Speech Recognition in Radiology. Ph.D. thesis, Simon Fraser University, 2006

15. Sistrom C: Conceptual approach for the design of radiology reporting interfaces: the talking template. J Digit Imaging 18(3):176–187, 2005

16. Caviedes JE, Cimino JJ: Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform 37:77–85, 2004

17. Shiffman S, Detmer WMS, Lane, CD, et al: A continuous-speech interface to a decision support system: I. Techniques to accommodate misrecognized input. J Am Med Inform Assoc 2:36–45, 1995