Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 128 (2018) 32–37

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Natural Language and Speech Processing, ICNLSP 2015

Automatic Speech Recognition Errors Detection and Correction: A Review

Rahhal Errattahi[a,*], Asmaa El Hannani[a], Hassan Ouahmane[a]

[a]*Laboratory of Information Technologies, National School of Applied Sciences, University of Chouaib Doukkali, El Jadida - Morocco*

Abstract

Even though Automatic Speech Recognition (ASR) has matured to the point of commercial applications, high error rate in some speech recognition domains remain as one of the main impediment factors to the wide adoption of speech technology, and especially for continuous large vocabulary speech recognition applications. The persistent presence of ASR errors have intensified the need to find alternative techniques to automatically detect and correct such errors. The correction of the transcription errors is very crucial not only to improve the speech recognition accuracy, but also to avoid the propagation of the errors to the subsequent language processing modules such as machine translation. In this paper, basic principles of ASR evaluation are first summarized, and then the state of the current ASR errors detection and correction research is reviewed. We focus on emerging techniques using word error rate metric.

Keywords: Automatic Speech Recognition; ASR Error Detection; ASR Error Correction; ASR evaluation;

1. Introduction

Automatic Speech Recognition (ASR) systems aims at converting a speech signal into a sequence of words either for text-based communication purposes or for device controlling. The purpose of evaluating ASR systems is to simulate human judgement of the performance of the systems in order to measure their usefulness and assess the remaining difficulties and especially when comparing systems. The standard metric of ASR evaluation is the Word Error Rate, which is defined as the proportion of word errors to words processed.

ASR has matured to the point of commercial applications by providing transcription with an acceptable level of performance which allows integration into many applications. In general, ASR systems are effective when the conditions are well controlled. Nevertheless, they are too dependent on the task being performed and the results are far from ideal, and especially for Large Vocabulary Continuous Speech Recognition (LVCSR) applications. This later still one of the most challenging tasks in the field, due to a number of factors, including poor articulation, variable

* Corresponding author. Tel.: +212-523-344-822 ; fax: +212-523-394-915.
E-mail address: errattahi.r@ucd.ac.ma

10.1016/j.procs.2018.03.005

speaking rate and high degree of acoustic variability caused by noise, side-speech, accents, sloppy pronunciation, hesitation, repetition, interruptions and channel mismatch, and/or distortions. To deal with all these problems, there has been a plethora of algorithms and technologies proposed by the scientific communities for all steps of LVCSR over the last decade: pre-processing, feature extraction, acoustic modeling, language modeling, decoding and result post-processing. Nevertheless LVCSR systems are not yet robust with error rates of up to 50% under certain conditions [21],[8].

The persistent presence of ASR errors motivates the attempt to find alternative techniques to assist users in correcting the transcription errors or to totally automate the correction process. Manual errors correction is often tedious and time consuming. Hence automatic detection and correction of ASR errors has become an important research area, not only for improving speech recognition accuracy but also for avoiding the propagation of the errors to the post recognition process (e.g. Machine translation and Human-Computer interaction). The aim is to be able to automatically detect, classify, and then partially or fully correct errors, regardless of the ASR system used. This can be very effective, and particularly when the ASR system is used as a black-box and the user does not have access to tune the features, the models or the decoder of the ASR system.

In the present paper we present an overview about ASR errors and the stat-of-the-art techniques for their detection and correction so as to provide a technological perspective and an appreciation of the fundamental progress that has been made in this field.

2. ASR evaluation

The performance of any ASR system is evaluated in function of the error rate. The aim of ASR evaluation is to provide a comparison criterion between different systems or techniques and to measure the performance and the progress on specific tasks based on errors statistics. There are two key areas related to ASR errors, the first one is the reference-recognised alignment which consist of finding the best word alignment between the reference and the automatic transcription and the second one is the evaluation metrics measuring the performance of the ASR systems.

2.1. Performance Factors

ASR performance is dependent upon many different factors that could be grouped in the following categories:

- **Speaker Variabilities**: Usually the acoustic model is obtained using a limited amount of speech data that characterizes the speakers at a given time and situation. However, the voice can change in time due to aging, illness, emotions, tiredness and potentially other factors. For these reasons, the acoustic model may not be representative of all speakers in all their potential states. Variabilities may not all be covered, which affect negatively the performance of the ASR systems.
- **Spoken Language Variabilities**: The spontaneous and accented speech and the high degree of pronunciation variation due to dialects, and co-articulation are known to be critical for ASR. Also, with large vocabulary, it becomes increasingly harder to find sufficient data to train the language models. Thus, subwords models are usually used instead of words models which severely degrade the performance of the recognition.
- **Mismatch Factors**: The mismatch in recording conditions between the training and testing is the main challenge for speech recognition, specially when the speech signal is acquired on telephone lines. Differences in the background noise, in the telephone handset, in the transmission channel and in the recording devices can, indeed, introduce variabilities over the recording and decrease the accuracy of the system.

2.2. Reference-Recognised Word Sequences Alignment

There are three types of errors that occur in speech recognition. First, Substitution; where a word in the reference word sequence is transcribed as a different word. Second, Deletion; where a word in the reference is completely missed in the automatic transcription. And finally, Insertion; where a word appears in the automatic transcription that has no correspondent in the reference word sequence.

A key practical issue with ASR evaluation metrics calculation is finding the word alignment between the reference and the automatic transcription, which constitute the first step in the evaluation procedure. In other words, the reference and recognised words get matched in order to decide which word have been deleted or inserted, and which reference-recognised string pairs have been aligned to each other, which may result in a hit or a substitution.

This is normally done by using the Viterbi Edit Distance [17] to efficiently select the reference and the recognised word sequence alignment for which the weighted error score is minimized. The Edit Distance usually aligns an identical weights (1 for the Levensthein distance) to all three, insertion, substitution and deletion. Yet, unified weights may present a doubt to choose the best path alignment in the case when we have different ones which have the same score.

To avoid this problem Morris et al. [12] suggest using different weights, such that substitution will be favoured than insertion and deletion. In general, it's recommended to put $W_I = W_D$, and $W_S < W_I + W_S$. Where W_I, W_S and W_D are respectively the weight of insertion, substitution, and deletion.

2.3. ASR Evaluation Metrics

According to McCowan et al. [11] an ideal ASR evaluation metric should be: (i) Direct; measure ASR component independently on the ASR application, (ii) Objective; the measure should be calculated in an automated manner,(iii) Interpretable; the absolute value of the measure must give an idea about the performance, and (iv) Modular; the evaluation measure should be general to allow thorough application-dependent analysis.

Word Error Rate (WER) is the most popular metric for ASR evaluation, it measures the percentage of incorrect words (Substitutions (S), Insertions (I), Deletions (D)) regarding the total number of words processed. It is defined as

$$WER = \frac{S + D + I}{N_1} = \frac{S + D + I}{H + S + D} \tag{1}$$

where I = total number of insertions, D = total number of deletions, S = total number of substitutions, H = total number of hits, and N_1 = total number of input words.

Despite of being the most commonly used, WER has many shortcomings [10]. First of all, WER is not a true percentage because it has no upper bound, so it doesn't tell you how good a system is, but only that one is better than another. Moreover, WER is not D/I symmetric, so in noisy conditions WER could exceed 100%, for the fact that it gives far more weight to insertions than to deletions.

The WER still effective for speech recognition where errors can be corrected by typing, such as, dictation. However, for almost any other type of speech recognition systems, where the goal is more than transcription, it is necessary to look for an alternative, or additional, evaluation framework.

Many researchers have proposed alternative measures to solve the evident limitations of WER. In [12] Andrew et al. introduced two information theoretic measures of word information communicated. The first one, named Relative Information Lost (RIL), is based on Mutual Information (I, or MI) [7], which measures the statistical dependence between the input words X and output words Y, and is calculated using the Shannon Entropy H as follow:

$$RIL = \frac{H(Y|X)}{H(Y)} \tag{2}$$

with

$$H(Y) = -\sum_{i=1}^{n} P(y_i) log P(y_i) \tag{3}$$

and

$$H(X|Y) = -\sum_{i,j} P(x_i, y_j) log P(x_i, y_j) \tag{4}$$

Nevertheless, the RIL still too far from an adequate performance metric, since it is not simple to apply and it measures zero error for any one-one mapping between I/O words, which does not respond to the criteria of an ideal

ASR evaluation metric. The second one, named Word Information Lost (WIL), is an approximation measure of RIL. But, unlike RIL, WIL is simple to apply because it is based only on HSDI counts, and is given as :

$$WIL = 1 - \frac{H^2}{(H+S+D)(H+S+I)} \tag{5}$$

Table 1, extracted from [12], present a comparison between WER and WIL metrics.

Table 1. WER and WIL do not always give the same ranking (X, Y and Z are arbitrary words)

Input	Output	H	S	D	I	%WER	%WIL
X	X	1	0	0	0	0	0
X	XXYY	1	0	0	3	300	75
XYX	XZ	1	1	1	0	67	83
X	Y	0	1	0	0	100	100
X	YZ	0	1	0	1	200	100

Other metrics, that are dependent on the domain application of the ASR output, have also been explored in order to evaluate how useful that output would be to humans. In [14], Nanjo et al. defined Weighted Keyword Error Rate (WKER) as an evaluation metric for keyword-based open-domain speech understanding. Favre et al. [5] proposed an alternative evaluation metric to WER for the decision audit task of meeting recordings.

3. ASR errors detection and correction techniques

To enhance the performance of imperfect ASR systems, the automatic detection and correction of the transcription errors can, in some cases, be the only choice. Particularly when the ASR system is used as a black-box and the user does not have access to the internals of the system or when the manual correction is not convenient or even impossible as in the case where the transcription is not the final goal of the system.

3.1. ASR errors detection

The goal of errors detection is to determine whether an error has occurred in the transcription using features generated from the ASR system, such as confidence scores [9], language model, and confusion network density. Those features will be used later to classify a hypothesis word to two possible classes, either a correct word or an error. There are two categories of research that addressed the subject of errors detection in ASR systems: the first one focused on features generated from the ASR decoder, such as confidence scores, linguistic information, and confusion networks, and the second one used additional features generated from hypothesized word sequence, such as n-grams, parts of speech, syntactic features, and semantic features.

3.1.1. Decoder based features

Zhou et al. [23] addressed the issue of errors detection in ASR, especially in Dictation Speech Recognition (DSR), by using data-mining techniques. This study consists of using three different data-mining classifiers, including Nave Bayes (NB), Neural Networks (NN), and Support Vector Machines (SVM) for detecting errors in DSR. The three models were trained to identify errors using features extracted from DSR output, including confidence scores and linguistic information. Results of this study have shown that those systems could identify until 50% of output errors.

Another study [2] proposed the use of additional features extracted from the confusion networks, and estimated a correctness probability using logistic regression based on those features. The proposed system achieved a classification error rate of 12.3% on a French broadcast news corpus.

Pellegrini et al. [16] investigated the use of a Markov Chains (MC) classifier with two states: error state and correct state, to model errors using a set of 15 common features in errors detection. The resulted system was tested on American English broadcast news speech NIST corpus, and has achieved 860 errors correctly detected with only 16.7% classification error rate.

Chen et al. [4] proposed a system for errors detection in conversational spoken languages translation, in addition to traditional features obtained from ASR outputs. This system used additional features provided as the feedback of Statistical Machine Translation (SMT), including SMT confidence estimates and posteriors from named entity detection (NED). Furthermore this system used an automated word boundary detector based on acoustic-prosodic features to verify the existence of ASR-hypothesized word boundaries, in order to improve the ASR errors detection. This system provided 2.8% absolute improvement in error detection over a simple error detector based on features traditionally employed in the literature (e.g. ASR confidence, LM perplexity, confusion network density, and phonetic acoustic model score deviation).

3.1.2. Non-decoder based features

Pellegrini et al. [15] suggested the use of non-decoder based features, extracted from other sources different than the ASR decoder, in addition to the traditional decoder based features. A binary word match feature that present a binary comparison between two different ASR systems, bigram hit feature measuring the number of hits found by querying a very popular Web search engine, and a topic feature to identify if a word is out of the global topic of the hypothesized sentence. The introduction of this non-decoder based features led to significant improvements, from 13.87% to 12.16% classification error rate with a maximum entropy model, and from 14.01% to 12.39% classification error rate with linear-chain conditional random fields, comparing to a baseline using only decoder-based features.

3.2. ASR errors correction

In general, automatic errors correction refers to the entire process including error detection. Since, correcting manually a large amount of data is often laborious and time consuming, there is a huge need to provide tools that could detect and correct automatically those errors. In other words, the aim is to be able either to detect and correct errors in the output of ASR systems without any human intervention.

To the best of our knowledge just few researches addressed the correction process of ASR errors, while the majority of researches were limited to the detection and suggested correcting erroneous segments manually.

Earlier researches focused on assisting users in the correction process by providing additional support. In [1, 13], the authors suggested exploiting alternative hypothesis generated by the ASR system to provide users with more choices in order to correct erroneous words. Others provided a navigation environment in order to help users to input there corrections to the computer [20, 6]. Another stream of researches consists of using user corrections to anticipate and correct other errors. Yu et al. [22] suggested to adapt the lexicon of an ASR system from the user's corrections. According to the authors, this method provided a WER reduction of 11% over the original output of the ASR system. Shi et al. [19] proposed the use of external informations, including word alternative hypothesis, noisy context and accurate context, to improve the performance of manual errors correction. The limitation of all these methods, however, is that they require human intervention.

Sarma et al. [18] build an ASR errors detector and corrector using co-occurrence analysis. They introduced a novel unsupervised approach for detecting and correcting miss-recognized query words in a document collection. According to the authors, this method can produce high-precision targeted detection and correction of miss-recognized query words. In the same context, Bassil and Semaan [3] proposed a post-editing ASR errors correction method based on Microsoft N-Gram dataset for detecting and correcting spelling errors generated by ASR systems. The detection process consists of detecting on-word spelling errors in reference with the Microsoft N-Gram dataset, and the correction process consists of two steps: the first one consists of generating correction suggestions for the detected word errors, and the second one, comprise a context-sensitive errors correction algorithm for selecting the best candidate for the correction. The error rate using the proposed method was around 2.4% on a dataset composed of a collection of five different English articles each with around 100 words read by five different speakers.

4. Conclusion

In this paper we presented a review of ASR errors detection and correction methods, putting an emphasis on the ones based on word error rate metric. There have been multiple researches, in the past 10-15 years, in improving the accuracy of ASR systems using the correction of the transcription errors. Even though the results are promising, the

majority of the researches are limited to the detection and suggest manual errors correction. Therefore, we believe there is a need of more investigation on automatic ASR errors correction and attention should be given to issues such as the efficiency, the usability and the robustness of the developed methods. Automating the ASR errors correction process can be very crucial, especially when tuning the ASR system itself is not allowed (e.g. the system is purchased as a black-box) or when the transcription is not the final goal of the system (e.g. machine translation).

Acknowledgements

This work is supported by the CNRST Morocco under Contract No. 5UCD2015.

References

[1] Ainsworth, W.A., Pratt, S., 1992. Feedback strategies for error correction in speech recognition systems. International Journal of Man-Machine Studies 36, 833–842.
[2] Allauzen, A., 2007. Error detection in confusion network., in: INTERSPEECH, pp. 1749–1752.
[3] Bassil, Y., Semaan, P., 2012. Asr context-sensitive error correction based on microsoft n-gram dataset. arXiv preprint arXiv:1203.5262 .
[4] Chen, W., Ananthakrishnan, S., Kumar, R., Prasad, R., Natarajan, P., 2013. Asr error detection in a conversational spoken language translation system, in: In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE. pp. 7418–7422.
[5] Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., et al., 2013. Automatic human utility evaluation of asr systems: does wer really predict performance?, in: INTERSPEECH, pp. 3463–3467.
[6] Feng, J., Sears, A., 2004. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. ACM Transactions on Computer-Human Interaction (TOCHI) 11, 329–356.
[7] HOFFMAN, J., 1967. Papoulis, a-probability random variables and stochastic processes.
[8] Jaitly, N., Nguyen, P., Senior, A., Vanhoucke, V., 2012. Application of pretrained deep neural networks to large vocabulary speech recognition, in: Proceedings of Interspeech.
[9] Jiang, H., 2005. Confidence measures for speech recognition: A survey. Speech communication 45, 455–470.
[10] Maier, V., 2002. Evaluating ril as basis for evaluating automated speech recognition devices and the consequences of using probabilistic string edit distance as input. 3rd year project, Sheffield University .
[11] McCowan, I.A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., Bourlard, H., 2004. On the use of information retrieval measures for speech recognition evaluation. Technical Report. IDIAP.
[12] Morris, A.C., Maier, V., Green, P., 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition., in: INTERSPEECH.
[13] Murray, A., Frankish, C., Jones, D., 1993. Data-entry by voice: Facilitating correction of misrecognitions, in: Interactive speech technology, Taylor & Francis, Inc.. pp. 137–144.
[14] Nanjo, H., Kawahara, T., 2005. A new asr evaluation measure and minimum bayes-risk decoding for open-domain speech understanding., in: In the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1053–1056.
[15] Pellegrini, T., Trancoso, I., 2010. Improving asr error detection with non-decoder based features., in: INTERSPEECH, pp. 1950–1953.
[16] Pellegrini, T., Trancoso, I., 2011. Error detection in broadcast news asr using markov chains, in: Human Language Technology. Challenges for Computer Science and Linguistics. Springer, pp. 59–69.
[17] Ristad, E.S., Yianilos, P.N., 1998. Learning string-edit distance. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20, 522–532.
[18] Sarma, A., Palmer, D.D., 2004. Context-based speech recognition error detection and correction, in: Proceedings of HLT-NAACL 2004: Short Papers, Association for Computational Linguistics. pp. 85–88.
[19] Shi, Y., Zhou, L., 2011. Supporting dictation speech recognition error correction: the impact of external information. Behaviour & Information Technology 30, 761–774.
[20] Suhm, B., Myers, B., Waibel, A., 2001. Multimodal error correction for speech user interfaces. ACM transactions on computer-human interaction (TOCHI) 8, 60–98.
[21] Swietojanski, P., Ghoshal, A., Renals, S., 2014. Convolutional neural networks for distant speech recognition. IEEE Signal Processing Letters 21, 1120–1124.
[22] Yu, D., Hwang, M.Y., Mau, P., Acero, A., Deng, L., 2004. Unsupervised learning from users' error correction in speech dictation., in: INTERSPEECH.
[23] Zhou, L., Shi, Y., Feng, J., Sears, A., 2005. Data mining for detecting errors in dictation speech recognition. Speech and Audio Processing, IEEE Transactions on 13, 681–688.