Proceedings of the International Conference on Spoken
Language Processing, Brisbane, Australia,
September 2008

An Empirical Analysis of Word Error Rate and Keyword Error Rate

Youngja Park[1], *Siddharth Patwardhan*[2], *Karthik Visweswariah*[3], *Stephen C. Gates*[1]

[1]IBM T.J. Watson Research Center, 19 Skyline Dr. Hawthorne, NY, USA
[2]School of Computing, University of Utah, Salt Lake City, UT, USA
[3]IBM India Research Lab, New Delhi, India
{young_park,scgates}@us.ibm.com, sidd@cs.utah.edu,v-karthik@in.ibm.com

Abstract

This paper studies the relationship between word error rate (WER) and keyword error rate (KER) in speech transcripts and their effect on the performance of speech analytics applications. Automatic speech recognition (ASR) systems are increasingly used as input for speech analytics, which raises the question of whether WER or KER is the more suitable performance metric for calibrating the ASR system. ASR systems are typically evaluated in terms of WER. Many speech analytics applications, however, rely on identifying keywords in the transcripts—thus their performance can be expected to be more sensitive to keyword errors than regular word errors.

To study this question, we conduct a case study using an experimental data set comprising 100 calls to a contact center. We first automatically extract domain-specific words from the manual transcription and use this set of words to calculate keyword error rates in the following experiments. We then generate call transcripts with the IBM Attila speech recognition system, using different training for each repetition to generate transcripts with a range of word error rates. The transcripts are then processed with two speech analytics applications, call section segmentation and topic categorization. The results show similar WER and KER in high-accuracy transcripts, but KER increases more rapidly than WER as the accuracy of the transcription deteriorates. Neither speech analytics application showed significant sensitivity to the increase in KER for low-accuracy transcripts. Thus this case study did not identify a significant difference between using WER and KER.

Index Terms: speech analytics, keyword error rate, word error rate

1. Introduction

Continuing advances in automatic speech recognition (ASR) have enabled many useful speech analytics applications such as topic identification and information retrieval of speech texts. The performance of speech recognition is typically measured using word error rate (WER), the ratio of word insertion, substitution, and deletion errors in a transcript to the total number of spoken words. While it is widely believed that word error rate of a speech recognition system strongly influences the performance of speech analytics systems [1, 2], some researchers noticed a divergence between word error rate and the accuracy of speech understanding systems [3, 4, 5, 6]. In certain cases, the accuracy of speech understanding applications improved even though WER in the speech transcripts increased [4, 6]. This has lead researchers to suggest that ASR systems for speech analytics should be trained to optimize keyword error rate (KER) instead of WER, since many speech analytics applications depend mostly on certain keywords. Sandness and Hetherington [5] propose a keyword-based method to train the acoustic model. Also, Nanjo and Kawahara [3] present a decoding strategy to minimize weighted keyword error rate.

The use of keyword error rate also raises the challenge of how to obtain the list of domain-specific keywords. Previous work has experimented with relatively constrained domains such as air travel as in ATIS (Air Travel Information System) [7] or weather in the JUPITER system [8], where keywords only included a very limited set of place names and expressions for date and time [5, 6]. Similarly, research by Nanjo and Kawahara [3] simply used all nouns, except pronouns and numbers, as keywords. Therefore, it still remains open questions how to best measure performance when ASR is used in combination with speech analytics, and whether WER or KER is the more suitable performance metric for calibrating the ASR system.

In this work, we address these questions by investigating the relationship between WER and KER and their effect on the performance of speech analytics applications for a more unconstrained domain of IT help desk contact center calls. The calls concern many different topics including issues with network connectivity, user password and various software tools. We first present an automatic method to extract domain-specific words from a given transcript collection, which can serve as keywords. We then use the IBM Research Attila Speech recognition toolkit to transcribe the calls [9], producing several transcripts with different level of accuracy by retraining the language model, and study the effect of WER and KER on two speech analytics applications: a topic categorization system that is highly keyword dependent, and a call section segmentation which does not rely on keywords.

Our experimental results show KER increases more rapidly than WER as the accuracy of speech recognition decreases. KER increases by 200% while WER only increases by 128%, from a transcript with 41.64% WER to a transcript with 18.24% WER. WER and KER are very similar for higher accuracy transcripts. The experiments did not reveal a significant effect of higher KER on the speech analytics applications.

2. Keyword Extraction and Keyword Error Rate Measurement

It is very difficult to define and collect keywords for an unconstrained domain such as contact center calls. In this work, we define keywords as domain-specific nouns and verbs. If a word occurs relatively more frequently in a domain-specific text than in a non-domain text, the word is regarded as domain-specific. Based on this notion, we define domain specificity of a word as the relative probability of occurrence of the word in a domain

text versus in a general text as shown in Equation 1.

$$domainspecificity(w) = \frac{p_d(w)}{p_g(w)} = \frac{\frac{c_d(w)}{N_d}}{\frac{c_g(w)}{N_g}} \quad (1)$$

where, $p_d(w)$ is the probability of word w in a domain-specific document, and $p_g(w)$ is the probability of word w in a general document collection. $c_d(w)$ denote the number of occurrences of word w in the domain text, and $c_g(w)$ is the number of occurrences of w in the general document collection. N_d and N_g are the number of words in the domain corpus and in the general corpus respectively. If word w does not exist in the general corpus, the count of w in the general corpus, $c_g(w)$, is set to 1 if w is a named entity such as abbreviations or proper names resulting in a high domain specificity. Otherwise, $c_g(w)$ is set to the highest count in the general corpus to generate a low domain specificity for w. In this work, $p_g(w)$ is computed based on a one million word corpus which consists mostly of news articles.

The extraction of domain-specific words is done by the following steps.

1. **Word normalization:** Word normalization is performed to aggregate all different expressions for a same concept. In speech transcripts, many words are used in several different variations such as inflections, abbreviations, and alternative spellings (e.g., UK and US spellings). We automatically identify all these variations of a word and aggregate them into a canonical form (see [10] for more detailed description).
2. **Identification of domain-specific words:** We compute the domain specificity values for all the canonical words using Equation 1, and select highly domain-specific nouns and verbs and their variations as keywords. The count of a canonical word is the combined count of all variations. Finally, we remove fillers (e.g., *um*, *ahh* and *yeah*), conversational contractions (e.g., *wanna*, *gonna* and *gotta*) and functional words from the keyword list.

We align each ASR transcript with its reference transcript, and compute KER based on the keyword set by using Equation 2.

$$KER = \frac{F + M}{N} \times 100 \quad (2)$$

where N is the number of keywords in the reference data, F is the number of falsely recognized keywords, M is the number of missed keywords.

3. Experimental Setup

In this section, we describe the selection of the experimental data and how to generate automatic transcripts with varying WER.

3.1. Experimental Data

The experiments are conducted with 100 calls to an IT help desk for a company. The 100 calls were randomly selected from 2,236 manually transcribed calls (about 300 hours) which were used to train the ASR system. The calls concern a variety of topics related to problems with network connectivity, user passwords and various software tools. The manual transcripts for the 100 calls are used as the reference data in this work, and thus both WER and KER of this data set are set to 0%. We call this data set $Transcript_0$ hereafter. Table 1 shows more information on the experimental data.

Number of calls	Total call length	Number of unique words	Number of word occurrences
100	13.8 hours	3,220	93,790

Table 1: Characterization of the experimental data.

3.2. Generation of ASR Transcripts with Varying WER

The ASR system uses a large US English vocabulary and works in speaker-independent mode. The acoustic model of the speech recognition system was trained on approximately 300 hours of 6kHz, mono audio data. For the speaker independent system, we used perceptual linear prediction (PLP) features, spliced and projected down using LDA+MLLT. The decision tree was trained with quinphone context, and we used about 50k Gaussians. The system uses vocal tract length normalization and feature space speaker adaptive training. We also use FMPE and MPE to train the speaker adaptive models. At test time, since both the speakers are on one channel, we use k-means clustering with a single Gaussian per speaker to determine speaker boundaries.

The ASR system exhibits 41.53% WER for the experimental data. To generate transcripts with improved accuracy, we interpolate our base language model with a language model built on the test data. We vary the interpolation weight to get varying error rates, and generated two sets of ASR transcripts with 18.23% and 26.75% respectively. Hereafter, the three sets of ASR transcripts are called $Transcript_{18}$, $Transcript_{27}$ and $Transcript_{42}$.

3.3. Generation of Keyword Set

We generate a keyword set from the manual transcripts (i.e., $Transcript_0$) using the method described in Section 2. The final keyword list contains 634 keywords, and they appear 7,729 times in $Transcript_0$. Note that the number of keywords comprises about 20% of the word types, but their occurrences amount only 8.25% of all the word occurrences. Table 2 displays 30 most domain-specific keywords extracted from $Transcript_0$.

Rank 1–10	Rank 11–20	Rank 21–30
um	ping	replicate
click	dos	scroll
yeah	*wanna*	yahoo
ah	ethernet	tivoli
server	navigator	icon
lan	folder	*gotta*
dot	reinstall	modem
password	adapter	sock
internet	thank	setup
com	database	explorer

Table 2: 30 most domain-specific words in the test data. Italicized words are fillers and conversational contractions, and excluded from the keyword list.

4. Speech Analytics Applications

In this section, we briefly describe the two speech analytics applications used for the analysis: a call segmentation system and a topic categorization system.

4.1. Call Section Segmentation

Many contact center calls follow a typical sequence during the conversation — greeting, question description, problem resolution and call closing. Identifying these segments in a call enables interesting applications for business insights and can improve search and retrieval functions for the call transcripts. We have built a Support Vector Machine (SVM)-based classification system which divides contact center call transcripts into a predefined set of call sections [11]. The target call sections include "Greeting Section", "Question Section", "Refine Section", "Research Section", "Resolution Section", "Closing Section" and "Out-of-topic Section".

The system first identifies utterance boundaries in a call transcript, and classifies each utterance into a call section. The classification is performed based on the following features.

- Speaker identification
- Call section type of the previous utterance
- Positions of the utterance in the transcript from the beginning and from the end of the call
- Number of domain-specific words in the utterance
- All content words that appear more than once in the training data

Note that the system uses many non-lexical features such as the speaker identification and positions of the utterance. The system, therefore, is expected to be not highly dependent on WER or KER of the speech recognition system.

For training and evaluation, we first manually mark the call sections in the 100 manual call transcripts. We then automatically align the manual transcripts with the call transcripts in $Transcript_{18}$, $Transcript_{27}$ and $Transcript_{42}$ using time stamps. This process produces automatic call transcripts annotated with call sections. We measure the performance of the call section segmentation system in terms of classification accuracy as shown below.

$$Accuracy = \frac{\text{number of correctly classified utterances}}{\text{total number of utterances in the test data}}$$

For evaluation on 100 calls, we conduct 10-fold cross validation and use the average classification accuracy for the study.

4.2. Topic Categorization

We have developed a keyword-based topic detection system which assigns one or more topics in a given taxonomy to a call. The system uses a human-generated domain taxonomy for the company. The taxonomy contains 158 distinct categories, and one or more keywords associated with each category. The topic categorization system finds the keywords defined in the taxonomy in a given call, and assigns the topics with which the keywords are associated. This system, therefore, is expected to be very sensitive to keyword recognition error of the ASR system.

We measure the performance of the topic categorizer in terms of precision and recall. Precision and recall are most widely used metrics in information retrieval and topic categorization [12]. For each call, precision and recall are calculated as defined in Equation (3) and 4.

$$precision = \frac{\text{number of correctly assigned topics}}{\text{number of topics assigned for the call}} \quad (3)$$

$$recall = \frac{\text{number of correctly assigned topics}}{\text{number of correct topics}} \quad (4)$$

In this work, we report the macro-average of precision and recall on the 100 calls. Macro-average precision(recall) is the mean of the precision(recall) values of all calls.

5. Experimental Results

In this section, we show the experimental results on the relationship between WER and KER, and their effect on the performance of the applications.

5.1. Relationship between WER and KER

Table 3 shows word error rate and keyword error rate of the four transcript sets. Note that KER is higher than WER in $Transcript_{27}$ and $Transcript_{42}$. We can also observe from Figure 1 that keyword error rate increases more rapidly than word error rate as the recognition accuracy deteriorates. KER increases 71% from $Transcript_{18}$ to $Transcript_{27}$ and 200% from $Transcript_{18}$ to $Transcript_{42}$ while WER increases 47% and 128% respectively.

The 'Ratio' column in the table shows the percentage of keyword errors over word errors. In the manual transcripts, 7,729 out of 93,790 word occurrences are keyword occurrences (i.e, 8.24% of word occurrences). Note that the ratio of keyword errors in $Transcript_{27}$ is 8.26%, which suggests that the ASR system misrecognizes keywords and non-keywords at almost same rate when the ASR has about 25% WER. When the ASR's recognition accuracy is lower than the point, it makes more mistakes for keywords.

Transcript Set	WER	KER	Ratio
$Transcript_0$	*0%*	*0%*	*0%*
$Transcript_{18}$	18.24%	16.28%	7.35 %
$Transcript_{27}$	26.77%	27.85%	8.26 %
$Transcript_{42}$	41.64%	48.84%	9.67%

Table 3: Word error rate and keyword error rate on the four transcript sets. WER and KER are very similar in high-accuracy transcripts, but KER is higher than WER in low-accuracy transcripts. The 'Ratio' column shows the percentage of keyword errors over word errors.

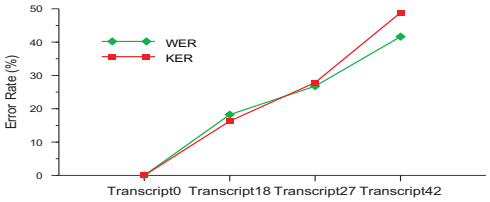

Figure 1: Changes of WER and KER across four transcript sets with varying WER. KER increases faster than WER as speech recognition accuracy deteriorates.

5.2. Effect of WER and KER on application performance

Table 4 summarizes the performance of the call section segmentation system and the topic categorization system. The call segmentation accuracy reported here is the average segmentation accuracy of 10-fold cross validation. Precision and recall of topic categorization are the macroaverage precision and recall on the 100 calls. The performance of the call section segmentation system decreases 7.5% from $Transcript_0$ to $Transcript_{42}$. The call segmentation shows the biggest performance drop (6.3%) from $Transcript_0$ to $Transcript_{18}$, and

Transcript Set	Call Segmentation Accuracy	Topic Categorization	
		Precision	Recall
$Transcript_0$	83.20%	32.34%	54.50%
$Transcript_{18}$	78.00%	31.90%	52.83%
$Transcript_{27}$	77.50%	25.22%	46.00%
$Transcript_{42}$	77.00%	23.14%	44.83%

Table 4: Performance evaluation results for the call section segmentation system and the topic categorization system.

a little difference across automatic speech transcripts. This can be explained by the fact that the call section segmentation system depends on speaker identification and utterance boundaries in the transcript, which are more accurate in manual transcripts than automatic speech transcripts.

We experience more performance degradation in the topic categorization system. Precision and recall fall 28.4% and 17.7% respectively from $Transcript_0$ to $Transcript_{42}$. We can also observe that precision is more vulnerable to recognition error than recall. The most performance degradation for the topic categorization occurs between $Transcript_{18}$ and $Transcript_{27}$, 21% and 13% for precision and recall respectively. Note that both precision and recall show little performance difference between $Transcript_0$ and $Transcript_{18}$, and between $Transcript_{27}$ and $Transcript_{42}$.

Finally, Figure 2 depicts the correlation of the two applications with WER and KER. The correlation coefficient, R, is measured by the following equation.

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y} \quad (5)$$

where n is the number of data samples, and μ_x (μ_y) and σ_x (σ_y) denote the mean and the standard deviation respectively for the variable x (y). As we can see from the figure, there is no significant difference in the correlation strengths between the applications' performance with WER and with KER.

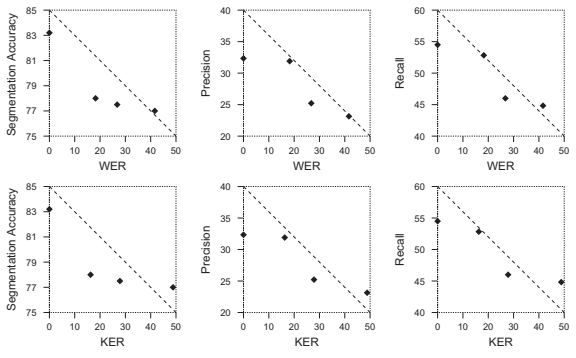

Figure 2: The performance of the call segmentation system and the topic categorization system as a function of WER and KER.

6. Conclusion

In this paper, we investigated the relationship between WER and KER in speech transcripts, and their effect on the performance of speech analytics applications. We also presented a method for automatically identifying domain-specific keywords.

The experiments were conducted with 100 calls to an IT help desk and two speech analysis applications: a call section segmentation system and a topic categorization system. We evaluated the applications using the manual transcripts of the calls and three sets of automatic call transcripts with varying WER.

Our experimental results show similar values for WER and KER in higher accuracy transcripts, but KER increases more rapidly than WER as the transcription accuracy deteriorates. The study did not reveal a significantly increased sensitivity of the speech analytics applications to higher keyword error rates. This suggest that the use of word error rate is sufficient especially for cases where WER remains below 25%.

7. Acknowledgements

The authors are very grateful to Chalapathy Neti and Wilfried Teiken for providing many insightful comments on this study. We also thank Keh-Shin Cheng for making the topic categorization system available for the experiment.

8. References

[1] M. Cavazza, "An empirical study of speech recognition errors in a task-oriented dialogue system," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001, pp. 1–8.

[2] G. Saon, B. Ramabhadran, and G. Zweig, "On the effect of word error rate on automated quality monitoring," in *Proceedings of Spoken Language Technology Workshop*, 2006, pp. 106 – 109.

[3] H. Nanjo and T. Kawahara, "A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding," in *Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, pp. 1053–1056.

[4] G. Riccardi, A. Gorin, A. Ljolje, and M. Riley, "A spoken language system for automated call routing," in *Proceedings of Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, 1997, pp. 1143–1146.

[5] E. D. Sandness and I. L. Hetherington, "Keyword-based discriminative training of acoustic models," in *Proceedings of ICSLP*, 2000, pp. 135–138.

[6] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 577–582.

[7] P. Price, "Evaluation of spoken language system: the ATIS domain," in *Proceedings of DARPA Speech and Natural Language Workshop*, 1990, pp. 91–95.

[8] V. Zue, S. Seneff, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, pp. 85–96, 2000.

[9] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 205–208.

[10] Y. Park, R. J. Byrd, and B. K. Boguraev, "Automatic glossary extraction: Beyond terminology identification," in *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING02)*, 2002, pp. 772–778.

[11] Y. Park, "Automatic call section segmentation for contact-center calls," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 117–126.

[12] D. Lewis, "Evaluating text categorization," in *Proceedings of Speech and Natural Language Workshop*, 1991, pp. 312–318.