Data Quality Relevance in Linguistic Analysis: The Impact of Transcription Error on Multiple Methods of Linguistic Analysis

Emergent Research Forum (ERF)s

Steven J. Pentland
Boise State University
stevenpentland@boisestate.edu

Lee A. Spitzley
University at Albany, SUNY
lspitzley@albany.edu

Christie M. Fuller
Boise State University
christiefuller@boisestate.edu

Douglas P. Twitchell
Boise State University
dougtwitchell@boisestate.edu

Abstract

There is an enormous amount of recorded speech generated daily, and quickly transcribing and analyzing the text of this speech could have tremendous value to organizations and researchers. However, the speech transcription process has historically been laborious, expensive, and slow. Automatic speech recognition (ASR) tools have matured a great deal in the last decade and may be a suitable method to generate large scale, high quality transcriptions. These tools are are fast and economical, but generally produce errors at a much greater rate than human transcribers. It is unknown whether these errors matter when conducting psycholinguistic research. In this study, we will investigate the accuracy of earnings conference call transcripts produced by multiple tools and the impact of that transcription accuracy on the results of subsequent text mining analysis. While prior studies have focused on a single form of text mining, we will conduct three types of text analysis: bag-of-words based classification, lexicon-based classification and sentiment analysis. The results will show whether a different level of transcription quality is required for different types of text mining and the feasibility of using automated transcription services across a range of text mining applications.

Keywords

Automatic Speech Recognition, Text Mining, Bag of Words, Sentiment Analysis

Introduction

Most corporate data available today is unstructured data, including text (Taylor 2017). The samples used for text mining applications can have a variety of origins, ranging from audio and video recordings to emails and documents. Our focus in this study is those sources that must be transcribed into text prior to analysis. Many automated speech recognition (ASR) tools are available, though the quality of transcriptions varies. The goal of this research is to understand the impact that transcript quality has on various text mining applications.

In the last few years, many new and updated transcription tools have been released. These include Siri, Amazon Transcribe, IBM Watson, Google Speech to Text, CMU Sphinx, Dragon and Microsoft Azure. Their cost is much less than that of manual transcription, and they provide faster turnaround time, enabling near real-time analysis of text. The availability of these tools provides the potential to analyze text that otherwise might remain unused. It also makes the transcribed text available to a broader audience, even those with

quite limited budgets. However, these are only advantages if the efficacy of analysis done on automated transcription is comparable to that of analysis performed on manual transcription or it can be established that lower quality transcriptions do not impact analysis.

While some sources report transcription accuracy of up to 95%, rivaling that of human transcription (Glaser 2017; Saon 2017), this accuracy appears to vary widely across tools and sample. One recent study reported Word Error Rates ranging from 7.3% to 54.2% depending on the corpus and tool (Dernoncourt et al. 2018). Another study reported word error rates ranging from 0 to 83% (Kepuska and Bouhata, 2017). Given these widely varying rates, it is important to establish the quality of transcription possible for a given set of tools and context. Further, the transcription quality required to perform different types of analysis is unknown. Prior research has shown that a fair amount of error can be present without impacting text classifications (Agarwal et al. 2007). That research focused on the classification of a minimum of seven and a maximum of over one hundred predefined categories. The acceptable error for other types of text mining applications, such as behavioral analysis, remains unknown.

This *research-in-progress* paper outlines the motivation behind our research, and proposes a methodology for understanding the influence of transcript quality on text mining methods. While overall word error rate is perhaps the most important measure of transcription quality, we also intend to study the relative impact of common errors such as deletions, insertions, and substitutions on text mining results. At the conclusion of this research, we hope to address the following questions: What is the quality of transcription produced by ASR for non-optimal audio input? What types of errors predominate the transcriptions? How does this quality impact the result of subsequent linguistic analyses? Can statistical power be calculated based on error rate?

Background

In many cases, transcriptions of speech are publicly available (i.e. presidential debates, quarterly earnings calls, and customer service calls). In these situations, interested parties have invested time and resources to manually annotate speech. However, these transcriptions are generally associated with lengthy processing times and high costs. The desire for accurate, real-time transcriptions is realized by the multitude of automated services currently on the market. Automated real-time transcriptions could prove advantageous for many tasks. For instance, financial markets fluctuate during quarterly earnings calls based on human interpretation of content (Mayew et al. 2016). Real-time computer analysis of linguistic content has potential to detect subtle nuances in speech during earnings calls that could prove beneficial to investors. Other studies have discovered linguistic cues to mental illness (Howes, Purver, & McCabe, 2014). In telehealth type situations, real-time speech to text combined with linguistic analysis could help support clinical diagnoses. Overall, automated transcription offers efficiency and scalability allowing greater access to linguistic data.

Text is increasingly being investigated from a behavioral perspective. Personality (O'Reilly et al. 2018), affect (Cheng et al. 2017), impression management (Pan et al. 2018), and veracity assessment (Zhou et al. 2004) are just a few constructs studied in relation to language. Arguably, distinguishing between extroversion/introversion, positive/negative affect, or truth/deception using message language is more complex than sifting and grouping message topics. Deception detection, for instance, is notoriously challenging with human accuracy rates little better than chance (DePaulo et al. 2003). Slight differences in language usage such as the use of first-person plural versus first-person singular can suggest characteristics like narcissism (O'Reilly et al. 2018) or deception (DePaulo et al. 2003). The subtleties of behavior specific cues and the sparseness of language in general highlights the potentially negative consequences of using imperfect datasets for linguistic behavioral inquiry.

Although linguistic analysis for behavioral understanding has received much attention, the work surrounding the impact of text errors has focused on topic modeling and information retrieval (e.g. Agarwal et al. 2007; Walker et al. 2010; Lopresti 2008). Agarwal et al. (2007) discovered that document noise as high as 40% had very little impact on supervised topic modeling. On the other hand, Walker et al.'s (2010) evaluation of non-supervised (clustering and LDA) methods did not yield the same robustness showing a negative correlation between noise and model performance. We have yet to realize the impact of transcript errors on text mining processes beyond topic modeling and information retrieval. It is likely that ASR will

not produce perfect transcription for quite some time. For those seeking to gain behavioral insights using text produced by ASR, it is crucial to understand the extent to which errors influence results.

Mediums such as web forums, instant messaging, SMS, emails, social media, and products reviews offer varying degrees and types of errors (Subramaniam et al. 2009). Emails, for instance, are relatively well composed, free of major errors or slang. Instant messages or SMS, on the other hand, are rife with errors and often littered with system specific nomenclature. Common text errors include spelling mistakes, deletion of characters, phonetic substitution, abbreviation, dialectal and informal usage, and deletion of words (Subramaniam et al., 2009). We categorize text errors as primary or secondary. Primary errors relate to human behaviors through mistakes or system specific nomenclature. SMS type platforms reflect high primary error environments where messages are produced and sent rapidly, many times with acronyms (i.e. lol, brb) and spelling errors. Secondary error generation is the result of converting written or spoken words to digital formats using tools such as optical character recognition (OCR) or automated speech recognition (ASR). In this case, written or spoken characters or words are misidentified by the tools. OCR technology commonly creates errors at the character level (Nagy et al. 1999) whereas ASR creates errors at the word level since the tools use predefined libraries when deciphering speech (Errattahi et al. 2018).

Methodology

To investigate the effects of transcription inaccuracies on behavioral text mining, this research will focus on the relationship between language during quarterly earnings calls and earnings surprises. During earnings calls, management provides financial updates to analysts and investors in a conference call setting. Earnings surprise refers to the difference between projected and actual earnings and is found to correlate with linguistic tone during calls. For instance, very positive (negative) linguistic tone predicts abnormally positive (negative) returns following the earnings call, beyond what is conveyed through financial information alone (Price et al. 2012). Abnormal positive tone is also associated with meeting or beating analyst forecasts (Huang et al. 2014). From a practical perspective, the verbalized content of earnings calls provides information beyond the explicitly communicated message. For example, language usage is found to correlate with future performance (Davis et al. 2006), as well as fraudulent actions within a company (Burgoon, et al. 2016; Larcker and Zakolyukina, 2012). The ability to transcribe and mine earning calls in near real-time using ASR provides an additional signal for making financial decisions in high-frequency trading environments.

From a research perspective, there are several benefits to using financial earnings calls for this current investigation. There is ample evidence from prior work to support relationships between earnings surprises, earnings call language, and economic consequences. Most publicly traded companies hold earnings calls each quarter, helping to create an extensive dataset. Text transcriptions of the calls are manually generated, and both transcription and original audio are publicly available. For investigatory purposes, this provides ground truth transcriptions, as well as audio for ASR processing. Financial earnings call data is similar to data in other domains, like telephone-based customer support or business meetings where transcripts are desired. In addition, earnings surprises are well-documented and lead to swift market corrections.

Sample

We will use quarterly earnings call transcripts from the financial news and information website Seeking Alpha (SA; seekingalpha.com) to conduct our analyses. SA transcribes, verifies, and posts earnings call transcripts within hours or days of an earnings call event. There are several important strengths of using this data. It is publicly available as SA posts new earnings calls without a paywall. Since approximately November of 2018, these calls have included an mp3 file containing the earnings call audio along with the transcript. This is important for reproducibility.

We estimate a sample of approximately 2500 calls based on an informal scan of SA, which revealed that SA posts about 85 calls per day, with approximately two-thirds of these calls being from US publicly-traded firms and containing the associated audio file. We intend to process about 40 business days (about two months) of calls. Earnings surprises will be implemented as a categorical variable--either negative or positive surprise if the reported earnings miss or beat analyst forecasts (Lu et al. 2018).

Planned Text Analysis

We intend to perform three types of text analysis: bag of words-based classification, lexicon-based classification and sentiment analysis and study the impact of transcription errors on each. We will use SpaCy (spacy.io) to perform the initial tokenization and sentence segmentation. SpaCy will also be used to create the TF-IDF weighting to be used for bag of words-based classification of our transcripts. For the lexicon-based classification of the transcripts, we will use SpaCy's matcher along with vocabularies from Loughran & McDonald (2011), LIWC (Pennebaker et al. 2001) and WordNet (Miller 1995) to build the transcript features for use as classification inputs. For both types of classification, the text features of the transcripts will be paired with the earning surprise target variable described above. For sentiment analysis, the sentiment will be generated by using the Loughran-McDonald Sentiment Word-Lists (Loughran and McDonald 2011).

For the current study, we will conduct two forms of validation: one with current period transcripts and results, to see if earnings surprises lead to differences in language used in the call (similar to Frankel et al. 2010, Matsumoto et al. 2011). We will also attempt to predict future earnings surprises or analyst forecast revisions following an earnings call.

Preliminary & Expected Results

Here we present both our preliminary results and describe the results that we expect in the completed version of this research. The preliminary results establish a baseline for ASR tools in our chosen context. The expected results will detail the full analysis of our corpus using multiple text mining approaches.

Transcription Quality

We used the method and software from Dernoncourt et al. (2018) to establish a baseline expectation for transcription quality. The current setup uses the IBM Watson Speech-to-Text API. This tool is designed to transcribe the audio from telephone calls and should be well-suited to the earnings call setting. We analyzed 19 conference calls that occurred in February 2019 to get an initial assessment of an ASR's quality on this type of recording. As shown in Table 1, the transcriptions contain errors at a far greater rate than would be expected from human transcription (~3% to 5%; Glaser 2017), and is much higher than the IBM's reported 5.5% WER (https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/). Our rates are similar to those found in other papers (e.g. Dernoncourt et al., 2018, Kepuska and Bouhata, 2017), and suggest that the error rate in the transcription varies with the audio source. As this research moves forward, we will use additional ASR tools to produce transcripts and will analyze the associated error rates for these tools.

Word Error Rate	Insertions	Deletions	Substitutions
34.51%	9.29%	6.17%	19.05%

Table 1. Transcription Quality for Preliminary Sample

Expected Results: Impact of errors on text analysis

To assess the impact of transcription errors on text analysis, we will compare the results of performing the analysis on both the transcriptions with errors and the error-free transcriptions across three types of text analysis. We expect transcription accuracy to lower the quality of predictions from our models. Under the assumption that the errors will in fact impact the results, we will also systematically adjust the level of error to determine at what point errors impacts text mining results. As is standard in classification studies, we will present measures such as overall accuracy, sensitivity, specificity, etc. (Japcowicz & Shaw, Chapter 2). We also anticipate analyzing the relative impact each type of error has on results.

Conclusion

In this study, we will investigate the quality of ASR tools for transcribing conference call recordings. Using the IBM Watson Speech-to-text tool on a small preliminary sample of 19 transcriptions, we found the WER is 34.51%, much greater than both the minimum reported WER of similar tools and the gold standard accuracy of human transcription. Moving forward, we will compare the accuracy of transcriptions with errors to that of perfect transcriptions across multiple types of text analysis, including bag of words-based classification, lexicon-based classification and sentiment analysis. These results will show whether perfect transcriptions are necessary for these tasks or whether reduced quality transcriptions can be substituted in any or all the studied analysis methods. If transcriptions with errors can be used, this opens the possibility of analysis of a wide set of text by many different researchers.

References

Agarwal, S., Godbole, S. Punjani, D., & Roy, S. 2007. "How Much Noise is Too Much: A Study in Automatic Text Classification," in Proceedings of the Seventh IEEE International Conference on Data mining, G. Jagannathan and R.N. Wright (eds.), Omaha, NE, pp. 3-12.

Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn,B., Byrd, & Spitzley, L. 2016. "Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls," *Journal of Language and Social Psychology*, (35:2), pp.123–157.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. 2017, February. "Anyone can Become a Troll: Causes of Trolling Behavior in Online Discussions," in *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work,* Portland, Oregon, pp. 1217-1230.

Davis, A.K., Piger, J.M., & Sedor, L.M. 2006. "Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases," No. 2006-005. Federal Reserve Bank of St. Louis.

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., & Cooper, H. 2003. "Cues to Deception," *Psychological Bulletin* (129:1), pp. 74–118.

Dernoncourt, F., Bui, T. & Chang, W. "A Framework for Speech Recognition Benchmarking," in *Proceedings of Interspeech 2018*, B. Yegnanarayana (chair), Hyderabad, pp. 169-170.

Errattahi, R., El Hannani, A., & Ouahmane, H. 2018. "Automatic Speech Recognition Errors Detection and Correction: A Review," *Procedia Computer Science* (128), pp. 32-37.

Frankel, R., Mayew, W. J., & Sun, Y. 2010. "Do pennies matter? Investor relations consequences of small negative earnings surprises". *Review of Accounting Studies*, 15(1), 220-242.

Glaser, A. 2017. "Google's ability to understand language is nearly equivalent to humans," https://www.recode.net/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

Howes, C., Purver, M., & McCabe, R. 2014. "Linguistic indicators of severity and progress in online text-based therapy for depression". *In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 7-16.

Huang, X., Teoh, S. H., & Zhang, Y. 2013. "Tone management". *The Accounting Review*, 89(3), 1083-1113.

Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge University Press.

Kepuska, V. & Bohouta, G. 2017. "Comparing Speech Recognition Systems (Microsoft API, Google API and CMU Sphinx," *International Journal of Engineering Research and Application* (7:3), pp. 20-24.

Larcker, D.F., & Zakolyukina, A.A. 2012. "Detecting Deceptive Discussions in Conference Calls," *Journal of Accounting Research* (50:2), pp.495-540.

Lopresti, D. 2009. "Optical character recognition errors and their effects on natural language processing". *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3), 141-151.

Loughran, T., & McDonald, B. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries and 10-Ks," *The Journal of Finance* (66:1), pp. 35-65.

Lu, R., Hou, W.X., Oppenheimer, H., & Zhang, T. (2018). "The Integrity of Financial Analysts: Evidence from Asymmetric Responses to Earnings Surprises," Journal of Business Ethics (151: 3), pp. 761-783.

Matsumoto, D., Pronk, M., & Roelofsen, E. 2011. "What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions". *The Accounting Review*, 86(4), 1383-1414.

Mayew, W.J., Sethuraman, M., & Venkatachalam, M. 2016. "'Casting' a Doubt: Informational Role of Analyst Participation during Earnings Conference Calls,"

Miller, G. A. 1995. "WordNet: A Lexical Database for English," *Communications of the ACM* (38:11), pp. 39–41.

Nagy, G., Nartker, T. A., & Rice, S. V. 1999. "Optical character recognition: An illustrated guide to the frontier," in *Document Recognition and Retrieval VII,* International Society for Optics and Photonics, pp. 58-70.

O'Reilly III, C. A., Doerr, B., & Chatman, J. A. 2018. "'See You in Court': How CEO narcissism increases firms' vulnerability to lawsuits," *The Leadership Quarterly (*29:3), pp. 365-378.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. 2012. "Earnings conference calls and stock returns: The incremental informativeness of textual tone". *Journal of Banking & Finance*, 36(4), 992-1011.

Saon, G. 2017. "Reaching New Records in Speech Recognition," IBM, https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

Subramaniam, L. V., Roy, S., Faruquie, T. A., & Negi, S. 2009, July. A survey of types of text noise and techniques to handle noisy text. *In Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. pp. 115-122. ACM.

Taylor, C. 2017. "Unstructured Data," Datamation, https://www.datamation.com/big-data/unstructured-data.html

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. 2004. "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications," Group decision and negotiation (13:1), pp. 81-106.