Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users

Akhter Al Amin, Saad Hassan, Cecilia O. Alm, Matt Huenerfauth
Rochester Institute of Technology
1 Lomb Memorial Drive, Rochester, NY
{aa7510,sh2513,coagla,matt.huenerfauth}@rit.edu

Abstract

Deaf and hard of hearing individuals regularly rely on captioning while watching live TV. Live TV captioning is evaluated by regulatory agencies using various caption evaluation metrics. However, caption evaluation metrics are often not informed by preferences of DHH users or how meaningful the captions are. There is a need to construct caption evaluation metrics that take the relative importance of words in a transcript into account. We conducted correlation analysis between two types of word embeddings and human-annotated labeled word-importance scores in existing corpus. We found that normalized contextualized word embeddings generated using BERT correlated better with manually annotated importance scores than word2vec-based word embeddings. We make available a pairing of word embeddings and their human-annotated importance scores. We also provide proof-of-concept utility by training word importance models, achieving an F1-score of 0.57 in the 6-class word importance classification task.

1 Introduction

Over 360 million people worldwide are Deaf or Hard of Hearing (DHH) (Mitchell et al., 2006; Blanchfield et al., 2001). In the U.S. alone, over 15% people are DHH, and regularly rely on captioning while watching videos to perceive salient auditory information (Berke et al., 2019). To provide quality captioning services to this group, it is essential to monitor the quality of captioning regularly. Regulators, e.g., the Federal Communication Commission (FCC) in the U.S. (Commission, 2014) are entrusted with regularly checking the quality of caption transcription generated by different broadcasters. However, given the abundant production of captioned live TV broadcasts, caption evaluation is a tedious and costly task.

DHH viewers are often dissatisfied with the quality of captioning provided in live contexts, which provide less time for caption production than pre-recorded contexts (Amin et al., 2021b; Kushalnagar and Kushalnagar, 2018). If regulatory organizations that measure the quality of captions used quality metrics that better reflect the DHH users' preferences, DHH viewers' experience may improve.

Existing metrics used in transcription or captioning include Word Error Rate (WER) (Ali and Renals, 2018) or Number of Error in Recognition (NER) (Romero-Fresco and Martínez Pérez, 2015). As noted by Kafle et al. (2019b), a major shortcoming of these metrics is that they do not consider the importance of individual words when measuring the accuracy of captioned transcripts (comparing to the reference transcript) and most metrics assign equal weights to each word. DHH viewers rely more heavily on important keywords while skimming through caption text (Kafle et al., 2019b).

Motivated by these shortcomings, prior work had proposed metrics which assign differential importance weights to individual words in captioned text when calculating an evaluation score (Kafle and Huenerfauth, 2019; Kafle et al., 2019a). Specifically, this prior work leveraged word2vec-based word embeddings to generate and propagate features to another layer of the network (Kafle and Huenerfauth, 2018). We build on this prior work and propose an updated approach. The feature space we are using contains both contextual and semantic information of the captioned text, which is crucial in conversational setting, often common in TV, and may better capture long-distance semantic and syntactic relationships. Thus, in this work, we contribute more current strategies for calculating importance of words in transcript text, toward a metric that takes word-importance into account when evaluating captions. Our contributions in this paper include:

1. **We conducted a comparative correlation analysis between human-annotated impor-**

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 35 - 40

tance scores for words in conversational transcripts and aggregated lexical semantic score generated from: (a) word2vec-based word embeddings as in prior work contrasted with (b) BERT-based contextualized embeddings. Our findings revealed that scores generated from contextualized embeddings had higher correlation with the human-annotated word-importance scores.

2. **We contribute data consisting of BERT contextualized word embeddings, paired with their word-importance scores, to augment a prior dataset of human-assigned importance scores for words in conversational transcripts (Kafle and Huenerfauth, 2018).** This enhanced data can be used by researchers for constructing improved caption-evaluation metrics or by researchers studying conversational discourse.

3. **To illustrate the use of this dataset, we show how interpretable classical machine-learning models can be trained to determine the importance of words using these contextualized word embedding vectors from our data.** In this proof-of-concept study, we show how these data can be used in training models. We leave detailed evaluation and comparison of models for future work.

2 Related Work

2.1 Word Importance Prediction

NLP researchers have explored approaches to determine word-importance for various downstream tasks, e.g. term weight determination when querying text (Dai and Callan, 2020), for text summarization (Hong and Nenkova, 2014) or text classification (Sheikh et al., 2016). Prior research on identifying and scoring important words in a text has largely focused on the task of keyword or important-term extraction (Dai and Callan, 2020; Sheikh et al., 2016). This task involves identifying words in a document that densely summarize it. Several automatic keyword-extraction techniques have been investigated, including unsupervised methods such as interpolation of Term Frequency and Inverse Document Frequency (TF-IDF) weighting (Sammut and Webb, 2010), Positive Pointwise Mutual Information (PPMI) (Bouma, 2009), word2vec embedding (Sheikh et al., 2016), and supervised methods that leverage linguistic features from text for word importance estimation (Dai and Callan, 2020; Kafle and Huenerfauth, 2018). While the conceptualization of word importance as a keyword-extraction problem has enabled retrieving relevant information from large textual or multimedia datasets (Dai and Callan, 2020; Shah and Bhattacharyya), this approach may not generalize across domains and functional, situational contexts of language use. For instance, given the meandering nature of topic transitions in television news broadcasts or talk shows (Kafle and Huenerfauth, 2019), when processing caption transcripts, a model of word importance that is more local may be more successful, rather than considering the entire transcript of the broadcast or show.

2.2 Caption Evaluation Methods

Several caption evaluation approaches have been proposed (Ali and Renals, 2018; Apone et al., 2011), with some approaches specifically taking into account the perspective of DHH participants (Kafle and Huenerfauth, 2018; Amin et al., 2021b). The most common caption evaluation used by different regulatory organizations is Word Error Rate (WER) (Ali and Renals, 2018). While penalizing insertion, deletion, and substitution errors in transcripts, a limitation of WER is that it considers importance of each word token equally. To address this, Apone et al. (2011) proposed a metric that assign weights to words in a text, but this probabilistic approach has not been trained on weights set to address priorities assigned by actual caption users.

In the most closely related work, Kafle and Huenerfauth (2018) investigated models for predicting word-importance during captioned one-on-one conversations. Their Automatic Caption Evaluation (ACE) framework utilized a variety of linguistic features to predict which words in a caption text were most important to its meaning, and which would be most problematic if incorrectly transcribed in a caption. Prior research on determining the importance of a word in a document had shown that an embedding can characterize a word's syntactic (e.g., word dependencies) and semantic character (e.g., named entity labeling), which in turn can help estimate a word's importance (Sheikh et al., 2016). Thus, Kafle and Huenerfauth (2018) used word2vec embeddings of words in the transcript. In this paper, we examine whether an alter-

native embedding, based on BERT, would lead to superior models of word-importance.

2.3 Annotation of Word Importance Scores

In this work, we contribute a dataset that augments a previously-released dataset from Kafle and Huenerfauth (2018), consisting of a 25,000-token subset of the Switchboard corpus of conversational transcripts (Godfrey et al., 1992). Kafle and Huenerfauth (2018) asked a pair of human annotators to assign word-importance scores to each word within these transcripts, on a range from 0.0 to 1.0, where 1.0 was most important. After partitioning scores into 6 discrete categories: [0-0.1), [0.1-0.3), [0.3-0.5), [0.5-0.7), [0.7-0.9), and [0.9 - 1], they trained a Neural Network-based classifier, using Long Short Term Memory (LSTM), to predict the importance category of each word in these transcripts. We augment this annotated corpus with recent contextualized word embeddings from BERT (Devlin et al., 2019), pairing up the embeddings with the hand-annotated word importance data.

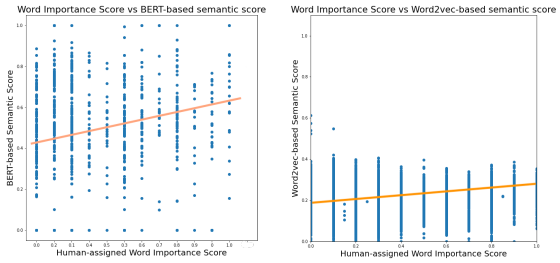
3 Corpus Augmentation

3.1 Extracting Word Embeddings Vectors

We have augmented the dataset described above, and will be releasing the version that includes two embeddings per word token: BERT contextualized word embeddings and word2vec embeddings. With this paper, we will be releasing the BERT-generated contextualized word embeddings[1] of 25,000 tokens, each with a feature vector of length 768, augmented with the human-annotated word-importance scores[2].

To enable comparison with the work of Kafle and Huenerfauth (2018), we extracted a word2vec (Rehurek and Sojka, 2011) embedding vector of length 100 for each word that occurred at least twice within each transcript. Next, we employed the pre-trained BERT model entitled *bert-base-uncased* (Devlin et al., 2019) to generate a contextualized word-embedding vector for each word within transcripts. For each word within each sentence, using BERT, we generated a three-dimensional embedding of shape $32 \times 12 \times 768$. These embeddings were created based upon the architecture of the pre-trained BERT model that included 32 transformer blocks, 12 attention heads and 768 hidden layers.

[1] https://nyu.databrary.org/volume/1447
[2] http://latlab.ist.rit.edu/lrec2018/



(a) (b)

Figure 1: Scatter plots for (a) the human-annotated score vs. BERT embedding-based semantic score, and (b) the human-annotated score vs. the word2vec embedding-based semantic score. The first 1200 words from the dataset are shown.

We follow prior work that has reshaped or composed the three dimensions into a one-dimensional vector while retaining similar semantic information (Turton et al., 2020). After performing these operations, for each word we obtained a contextualized embedding vector of length 768.

Method \ Word	*sunday*	*noise*	*plan*
Human-assigned score	0.60	0.40	0.70
BERT	0.10	0.42	0.61
word2vec	0.35	0.17	0.18

Table 1: Three sample words, *sunday*, *noise*, and *plan* have been excerpted from one transcript. The human-assigned importance of these importance score are 0.60, 0.40, and 0.70. For *noise* and *plan*, aggregated scores generated from word2vec-based embedding are 0.17 and 0.18, which does not belong to the same importance categories annotated. On the contrary, Bert-based embedding generates a score that aligns with human-assigned importance for *noise* and *plan*. However, for *sunday*, the word2vec-based semantic score is relatively closer to the actual importance score than BERT-based embedding. In fact, *sunday* appears as an isolated response to someone's question in transcript.

3.2 Correlation Analysis to Assess Fit with Word Importance Scores

After calculating two types of embeddings for each word in this dataset, we asked which one would be more useful within a model to predict word importance. Prior work on the state-of-art word-importance learning algorithm Neural Bag-of-Words (NBOW) has revealed that learning importance of words within a sentence is effective while using the mean of each word-embedding vector as a feature (Sheikh et al., 2016). Following this common practice for determining word importance (Kalchbrenner et al., 2014; Dai and Callan, 2020), we calculated the mean of each word-embedding vector, to represent its word semantic score (Sheikh

Method	F1 Score	RMSE
Multi-layer Perceptron	0.10	1.29
Random-Forest	0.25	1.02
Linear Support Vector	0.51	0.99
Logistic Regression	0.57	0.92

Table 2: Supervised classification performance showing macro-averaged F1 score and Root Mean Squared Error.

et al., 2016). For both the word2vec and BERT-based embeddings, for each sentence in the transcript, we normalized word-semantic scores within the sentence, to obtain a value in a [0,1] range for each word. BERT embeddings produce sub-word tokens for a complete word and to handle such a scenario we have computed the average of the sub-words to calculate the final composite semantic score.

After performing this operation across sentences in the transcripts, we conducted an analysis to determine which form of pre-trained embedding (word2vec or BERT) better correlated with human-produced annotations of word importance in the original dataset. The values based on word2vec were correlated with human annotations with a Pearson correlation coefficient of $r = 0.30$, and for the BERT-based scores, the coefficient was $r = 0.41$. A Fisher z-transformation (Upton and Cook, 2014) revealed that word semantic scores generated using BERT contextualized word embeddings were significantly better correlated ($z = -3.05, p < 0.001$) with human-assigned scores than word2vec counterparts. Based on these findings, we decided to use BERT contextualized embeddings in continued analysis.

We also tried another traditional approach called TF-IDF to calculate a semantic score for words. A correlation analysis between the score generated by TF-IDF and human annotations resulted in a Pearson correlation coefficient of $r = 0.25$, which was lower than the coefficient generated using word2vec word embedding.

4 Predicting Word Importance

To demonstrate how to use our dataset to predict the importance of each word, we have begun to investigate several supervised learning methods. The independent variable is the processed 768×1 BERT-embedding vector of each word, and the output variable is the human-labeled importance score, discretized into six classes, for each word in the dataset. This classification experiment partitioned the corpus into 80% training, 10% development, and 10% test set. This partition has been directly adapted from (Kafle and Huenerfauth, 2018). We evaluated the model using two measures: (i) Root Mean Square Error (RMSE) - the deviation of the model predictions from the human-assigned categories, and (ii) the F1 measure for classification performance. For classification, we categorized annotation scores into the 6 levels, as described above: [0-0.1), [0.1-0.3), [0.3-0.5), [0.5-0.7), [0.7-0.9), and [0.9 - 1].

	Predicted Label					
True Label	1	2	3	4	5	6
1	0.69	0.21	0.18	0.15	0.18	0.00
2	0.22	0.64	0.25	0.26	0.13	0.33
3	0.05	0.12	0.48	0.11	0.18	0.00
4	0.02	0.02	0.03	0.48	0.06	0.11
5	0.01	0.01	0.04	0.00	0.40	0.00
6	0.00	0.00	0.00	0.00	0.00	0.56

Table 3: Normalized confusion matrices for Logistic Regression for classification into six word importance classes using BERT-generated embeddings-based score.

Table 2 illustrates that the better performing supervised model (of four traditional approaches) in predicting the importance class is Logistic Regression with F1-score 0.57 and RMSE 0.92. Even if the classes are discretized, we are generating continuous value for each word. And since both the human and supervised model generated scores, we calculated this RMSE. Among other approaches, the Linear Support Vector Classifier achieves F1-score 0.51, Random-Forest achieves 0.25, and Multi-layer Perceptron achieves 0.10.

5 Limitations and Future Work

There are several limitations of this ongoing research that we intend to address in future work.

- In our current research, we have determined a semantic score for each word using three methods. Future research can use other methods to generate the semantic score and retrospectively compare the generated semantic score with the score assigned by the human annotators.
- The findings from this analysis leaves the room for future improvements, since we did not modify the hyperparameters to observe how accurately the models would predict the importance of words. Therefore, future research can explore variations of these models.
- Future directions may include collecting additional data to balance the distribution of im-

portance classes. In addition, given the role of part of speech (POS) for word importance in texts (Shah and Bhattacharyya), a next step could be to investigate POS with contextual word embedding for predicting word importance. Since TV captions often represent conversational speech with filler words, e.g., *hmm* or *yeah*, future research could consider alternative strategies to score the importance of such words.

- Hutchinson et al. (2020) and Hassan et al. (2021) demonstrate that a large language model like BERT can introduce bias relating to people with disabilities into a task. Therefore, future work can investigate whether BERT is introducing any latent bias in predicting importance of words from DHH viewers' perspective.

6 Conclusion

The analysis presented above has revealed that BERT contextualized word-embedding can better represent the importance of words compared to word2vec embeddings, which had been used in prior work on word-importance prediction (Kafle and Huenerfauth, 2019). Research indicates that DHH viewers often follow key terms while skimming through captions, and researchers have proposed approaches to guide DHH readers to quickly identify keywords in caption text through visual highlighting (Kafle et al., 2019b). Our findings may allow broadcasters to use embeddings to determine the important words within a sentence and to highlight those words in captions, to support DHH viewers' ability to read (Amin et al., 2021a) the captions effectively. In this study, a traditional Logistic Regression algorithm performed better at predicting importance classes.

We are also broadly investigating how to accurately measure the quality of caption transcriptions that are broadcast during live TV programs from the perspective of DHH viewers. We plan to incorporate predictive models into new word-importance weighted metrics, to better capture the usability of live captioning from DHH users' perspective.

7 Ethics Statement

This work advocates for improved inclusion of DHH individuals. A risk of the study is that results may not generalize across conversational corpora.

Acknowledgments

This material is based on work supported by the Department of Health and Human Services under Award No. 90DPCP0002-0100, and by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Health and Human Services or National Science Foundation.

References

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Akhter Al Amin, Abraham Glasser, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2021a. Preferences of deaf or hard of hearing users for live-TV caption appearance. In *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments*, pages 189–201, Cham. Springer International Publishing.

Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021b. Caption-occlusion severity judgments across live-television genres from deaf and hard-of-hearing viewers. In *Proceedings of the 18th International Web for All Conference*, W4A '21, New York, NY, USA. Association for Computing Machinery.

Tom Apone, Brooks Marcia Botkin, Brad, and Larry Goldberg. 2011. Caption accuracy metrics project research into automated error ranking of real-time captions in live television news programs.

Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred appearance of captions generated by automatic speech recognition for deaf and hard-of-hearing viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA. Association for Computing Machinery.

Bonnie B Blanchfield, Jacob J Feldman, and Jennifer L Dunbar. 2001. The severely to profoundly hearing-impaired population in the united states: prevalence estimates and demographics. *Journal of the American Academy of Audiology*, 12(4).

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tubingen.

Federal Communications Commission. 2014. *Closed Captioning of Video Programming; Telecommunications for the Deaf and Hard of Hearing, Inc. Declaratory Ruling, FNPRM*. Consumer and Governmental Affairs, Washington, D.C., USA.

Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*, WWW '20, page 1897–1907, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Sushant Kafle, Cecilia Ovesdotter Alm, and Matt Huenerfauth. 2019a. Fusion strategy for prosodic and lexical representations of word importance. In *Proc. Interspeech 2019*, pages 1313–1317.

Sushant Kafle and Matt Huenerfauth. 2018. A corpus for modeling word importance in spoken dialogue transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 99–103, Miyazaki, Japan. European Language Resources Association (ELRA).

Sushant Kafle and Matt Huenerfauth. 2019. Predicting the understandability of imperfect English captions for people who are deaf or hard of hearing. *ACM Trans. Access. Comput.*, 12(2).

Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019b. Evaluating the benefit of highlighting key words in captions for people who are deaf or hard of hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 43–55, New York, NY, USA. Association for Computing Machinery.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Raja Kushalnagar and Kesavan Kushalnagar. 2018. Subtitleformatter: Making subtitles easier to read for deaf and hard of hearing viewers on personal devices. In *Computers Helping People with Special Needs*, pages 211–219, Cham. Springer International Publishing.

Ross E Mitchell, Travas A Young, Bellamie Bachelda, and Michael A Karchmer. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Language Studies*, 6(3):306–335.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Pablo Romero-Fresco and Juan Martínez Pérez. 2015. *Accuracy Rate in Live Subtitling: The NER Model*. Audiovisual Translation in a Global Context. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, London.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Chirag Shah and Pushpak Bhattacharyya. A study for evaluating the importance of various parts of speech (pos) for information retrieval.

Imran Sheikh, Irina Illina, Dominique Fohr, and Georges Linares. 2016. Learning word importance with the neural bag-of-words model. In *ACL, Representation Learning for NLP (Repl4NLP) workshop*, Proceedings of ACL 2016, Berlin, Germany.

Jacob Turton, David Vinson, and Robert Elliott Smith. 2020. Deriving contextualised semantic features from bert (and other transformer model) embeddings.

G. Upton and I. Cook. 2014. *A Dictionary of Statistics 3e*. Oxford Paperback Reference. OUP Oxford.