A Performance Analysis of Word Importance Model for Predicting Importance of Words in Captions

Akhter Al Amin

December 9, 2021

1 Introduction

15% of US adults who are DHH rely on captioning while watching videos to perceive salient auditory information [1]. To support this special interest group with quality captioning, it is essential to provide accurate transcription. However, several reports have shown that the accuracy of transcription needs to be monitored to on a regular basis. Specially, regulators e.g., Federal Communication Commission(FCC) needs to regularly check the quality of caption transcription generated by different broadcasters. Given the abundant production of captioning, regulators tend to rely on automatic caption quality measurement technique. This method has increased the volume of measurement performed. However, there remains a concern that existing caption metrics consider the importance or each word equally which does not reflect DHH viewers' perspective. As prior research have show that DHH viewers tend to read keywords instead of reading the whole caption text to understand the context [2]. Therefore, current caption metrics are unable to measure caption transcription quality from DHH viewers' perspective. Considering the gap that exist in current literature, we need to understand whether we can measure the caption quality in a more inclusive way. In this work, we have explored X Natural Language Processing(NLP) algorithms to measure importance of wards for a given caption transcription and empirically evaluated how these algorithms performed in comparison with expert annotators' annotated word importance. Also we have compared the performance with state-of-the-art LSTM-based model proposed in [3].

2 Methods

2.1 Dataset

The Switchboard corpus consists of audio recordings of approximately 260 hours of speech consisting of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from across the United States. In January 2003, the Institute for Signal and Information Processing (ISIP) released written transcripts for the entire corpus, which consists of nearly 400,000 conversational turns. The ISIP transcripts include a complete lexicon list and automatic word alignment timing corresponding to the original audio files [3].

The importance score ranges from 0 to 1. However, the paper, referred above, have categorized the words based on the importance level as follows:

- Importance 1: [0-0.1)
- Importance 2: [0.1-0.3)
- Importance 3: [0.3-0.5)
- Importance 4: [0.5-0.7)
- Importance 5: [0.7-0.9)
- Importance 6: [0.9-1]

The higher the importance value is, the higher the weight of the word.

2.2 Methods

We have implemented several word importance methods to compare with the state-of-the-art LSTM-based approaches. Initially, we implemented several unsupervised approaches as follows:

1. Bag-of-Words (BOW)
2. Term frequency and Inverse Document Frequency (TF-IDF)
3. Word Embedding(CBOW) summation [4] with interpolation of Parts of Speech (POS) importance [5]
4. Composition of Contextualized Word Embedding (BERT)

After observing the performance of unsupervised approaches, we realized that it would interesting to see how supervised models perform which is trained on the small annotated dataset described in section 2.1. Therefore, we implemented supervised learning approach, wherein we have employed the Contextualized Word Embedding (BERT). We have adapted this method from a prior document-based word importance model proposed in [6]. After cross-validating on 4 supervised learning approaches (1) Multinomial Naiv-Bayes, (2) Logistic Regression (3) Linear Support Vector Classifier and (4) Random Forest Classifier, we have selected Logistic Regression as best performed model.

3 Results

The analysis we have conducted using different methods have revealed how well each of these approaches can predict importance of the words. We have compared the findings with the word-by-word importance corpora, described in section 2.1. It is important to remember that since there are 6 importance classes, we will represent the comparison of resulting Macro-Average Precision(MAP) for across 4 approaches.

Note that in this comparison, we are excluding BOW approaches, as in this approach we are unable to dissect the importance of word in six category. Since, BOW approach only provide importance of word either 0 or 1.

The table above illustrates the comparison of performance of different word-importance prediction methods.

Method	F1 score (Mean Average Precision)
TF-IDF	0.15
Word Embedding Summation * POS importance	0.26
Contextualized Word Embedding	0.19
Random-Forest Classifier	0.25
Logistic Regression	0.54
LSTM (State-of-the-art)	0.60

Table 1: Model performance in terms of macro-averaged F1 score, with best results and state-of-the-art in bold font.

4 Reflection

From the analysis, we are going illustrate some reflections below:

4.1 Best Performed Unsupervised Method

Among the unsupervised approaches, discussed above and shown in table 1, we can observe that interpolation of static word embedding and POS importance outperform other methods. Specially, the count-based method like BOW and document based word importance approach TF-IDF. The main reason of poor performance of count-based approach is inclusion of lots of filler words in stead of content and function words. We realized that using POS might be able to encounter the issue with filler words as that might allow us to put some extra weight on content and function words. And that is the primary reason why this approach have outperformed other approaches.

4.2 Feature Space

We have been able to use several feature for each of this approach. The most effective approach is using word embedding. It has be discussed earlier that Word Embedding contain a great deal of rich semantic and contextual information about the words. Previously a bi-directional word embedding was used to general and propagate features to next layer of the network [3]. Since the network involves a forget gate that decide which information forget and which to keep, there remains a possibility of forgetting essential semantic information. In current approach, the feature space we are using contains both contextual and semantic information of the document which is crucial in conversational setting. We argue that this approach may be able to resolve the problem related to long-distance semantic relationship.

4.3 Best Performed Supervised Method

We have tried several supervised approaches e.g. Logistic Regression, Random Forest Classifier, Linear Support Vector Classifier. After evaluating the accuracy of the cross-validation set of the data, we selected logistic regression as best performed model with F-score 0.54.

4.4 Comparison with State-of-the-art Neural Network-based Approach

Existing state-of-the-art LSTM-based approach achieved F1 score 0.60, whereas our best performed Logistic Regression-based model revealed 0.54. Also this approach did not outperformed existing method. But this approach performed better for predicting importance class 4 to importance class 6.

5 Improvement Scope

The future direction of this works are as follows:

- Neural Network-based approach with similar feature space might be interesting aspect look at. In this research, due to limitation of data, we have been unable to implement Neural Network based approaches and compare empirically.
- Since TV captions, largely, represent conversational text and in conversation, we have a tendency to use a lot of filler words, e.g. 'humm', 'yeah' which are not either content or function words, it is crucial to consider this aspect while collecting word importance annotation from annotators.
- We observed from all these approaches we have implemented above lot of them do not consider the process how DHH viewers extract information from a certain set of text. Existing unsupervised approaches which are mostly count-based do not consider the semantic definition or how much information a specific word carries given a specific context or document to DHH viewers. Also in supervised approaches that tend to employ word embedding which might include both syntactic and semantic definition which we can observe in word similarity analysis. However, the word embedding approaches, both static and dynamic, are not informed by DHH viewers' preference. A future work can investigate how to make this word embedding approaches more inclusive.
- Future research can investigate how interpolation of POS importance with BERT-based word embedding work in predicting word-importance unsupervised manner.

References

[1] L. Berke, K. Albusays, M. Seita, and M. Huenerfauth, "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, (New York, NY, USA), p. 1–6, Association for Computing Machinery, 2019.

[2] S. Kafle and M. Huenerfauth, "Predicting the understandability of imperfect english captions for people who are deaf or hard of hearing," *ACM Trans. Access. Comput.*, vol. 12, June 2019.

[3] S. Kafle and M. Huenerfauth, "A corpus for modeling word importance in spoken dialogue transcripts," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[4] I. Sheikh, I. Illina, D. Fohr, and G. Linares, “Learning Word Importance with the Neural Bag-of-Words Model,” in *ACL, Representation Learning for NLP (Repl4NLP) workshop*, Proceedings of ACL 2016, (Berlin, Germany), Aug. 2016.

[5] C. Shah and P. Bhattacharyya, “A study for evaluating the importance of various parts of speech (pos) for information retrieval,” in *IR). Proceedings of International Conference on Universal Knowledge and Languages (ICUKL) 2002*, 2002.

[6] Z. Dai and J. Callan, “Context-aware document term weighting for ad-hoc search,” in *Proceedings of The Web Conference 2020*, WWW ’20, (New York, NY, USA), p. 1897–1907, Association for Computing Machinery, 2020.