A comprehensive word difficulty index for L2 listening

Kourosh Meshgi, Maryam Sadat Mirzaei

Graduate School of Informatics, Kyoto University, Japan

<https://doi.org/10.36505/ExLing-2018/09/0019/000352>

Abstract

Word difficulty in the listening task is considered challenging because of high subjectivity, high dimensionality, and low generalizability. We propose a word listening difficulty score, as a linear combination of several complementary features. A dataset of expert-annotated partial and synchronized captions for TED talks is prepared for a target language proficiency, in which only the difficult words are shown. A linear SVM was trained on this dataset, and the learned parameters of the SVM were transferred to the proposed score. This data-driven score demonstrates higher accuracy on the annotated dataset and facilitates model and feature expansion.

Key words: Word listening difficulty score, partial and synchronized caption

Introduction

We have developed a tool, partial and synchronized caption (PSC), to foster L2 listening skill by providing an adequate amount of text in the caption to provide assistance to the listeners only when they encounter difficulties. PSC uses automatic speech recognition (ASR) technology to realize word-level text-to-speech alignment and makes a principled selection of words for inclusion in the caption (Fig. 1) based on word frequency, speech rate, and specificity. Our earlier studies using these features revealed that PSC leads to the same level of comprehension as the full-caption, but better prepares learners for listening without using any textual clues (Mirzaei et al., 2017).

To cover various sources of listening difficulties (Bloomfield et al., 2010), it is possible to incorporate more features to the PSC system: lexical features, acoustic/speaker features, content attributes, and perceptual hindering factors. However, these features may be correlated with each other and finding their footprint in listening difficulty is not straightforward. Furthermore, not all the features are equally useful for detecting difficult words in the speech. To gauge the difficulty of a word, readability scores (e.g., Dale-Chall score) consider the frequency of the words in corpus but discard the speaker/acoustic information. Yoon et al. (2016) proposed a listenability score, by fitting a regression model on the average of Likert-scale difficulty scores for each word, given by

several non-native speakers with different proficiencies. Their model cannot generalize well on out-of-vocabulary data, and their process was subjective and dependent on data obtained from non-experts. Kotani & Yoshimi (2017) further explored this idea by using learner proficiency as a feature in the score.

Figure 1. Screen shot for a TED talk with word-level synchronized caption (left) and partial and synchronized caption (right).

This study approaches this problem by extracting various features from speech and find the degree by which they contribute to listening difficulty of each word for a certain proficiency of L2 learners. Using this importance weights, a data-driven word difficulty index is proposed to predict the difficulty of each word in an L2 listening task for a target proficiency.

Methodology

Selecting the difficult words for a target language proficiency can be reformulated as a binary classification problem where, a classifier defined by model θ tries to score word w_i through the use of classifier confidence function $h(w_i | \theta): \{\text{vocabulary}\} \rightarrow [0,1]$. This study strives to define such function as a linear combination of given features f_j :

$$h(w_i | \theta) = \sum_{j=1}^n \lambda_j K(f_j(w_{i-\Delta i+\Delta'}), \theta_j) \quad (1)$$

Where $K(\cdot)$ is a kernel function and λ_j denotes the weight of the j -th feature. Such a scoring function may measure word listening difficulty and can be incorporated in PSC. In this view, the words of a speech's transcript are scored, and sorted based on their difficulty. Then the desired number of difficult words are selected from this list to be shown in the caption, which enables adjusting the amount of words flexibly.

Dataset: We prepared a dataset of several TED talks (~ 90 min including $\sim 10,200$ words) given by American native speakers, forced aligned them using Kaldi ASR and Gentle forced aligner (*v0.10.1*), and provided them to two experts to be labeled for intermediate English

proficiency. The labels are *show* (for difficult words) and *hide* (for others). Both annotators had linguistic backgrounds and received a set of clear instructions and objectives, that resulted in a high inter-annotation agreement (Cohen’s $\kappa = 0.83$).

Features: Figure 2 shows a RadViz chart of most informative features (explained later) extracted from the annotated dataset. RadViz projects a high-dimensional feature set into a 2D space where the influence of each feature can be shown as a balance between all of the features.

Figure 2. RadViz graph of most informative features extracted from the dataset. Figure 3. Accuracy of different classifiers, decision tree (DT), naïve Bayes (NB) and linear support vector machine (SVM).

Classifiers: We conducted an experiment to assess the performance of different classifiers on this classification task given the obtained human annotation as ground truth. To this end, a J48 decision tree (DT), a naïve Bayes classifier (NB), and a linear support vector machine (SVM) is trained on the annotated data via a 5-fold cross-validation. Figure 3 shows that while NB and SVM demonstrate comparable accuracy and F1-score. When generating PSC, however, recall is more important than the precision. Precision indicates the portion of shown words that are actually difficult, whereas recall indicates the portion of the difficult words that are shown. Therefore, SVM that shows more of difficult words in the PSC is preferred. Nevertheless, NB helped to find the most informative features (Fig. 2).

Scoring Function: The superior performance of linear SVM in the given classification task, inspired us to create a word listening difficulty score as a linear combination of features where the coefficient of different features is transferred from the slope of the SVM’s decision boundary ($\lambda_j^{(SVM)}$), and the magnitude of the feature is calculated as its distance from the decision boundary in that specific dimension ($\theta_j^{(SVM)}$).

$$h(w_i | \theta) = \sum_{j=1}^n \lambda_j^{(SVM)} |f_j(w_{i-\Delta:i+\Delta'}) - \theta_j^{(SVM)}| \quad (2)$$

It should be noted that any arbitrary kernels may be considered to calculate this score, as it will be shown in the next section.

Evaluation

We generated partial and synchronized captions for the annotated TED talks using the proposed score function (*Score-L*), an improved version of the score using Gaussian kernel (*Score-K*) and expert rules (*PSC1*) described in Mirzaei et al. (2017) for intermediate L2 learners. The number of shown words in PSCs using proposed scores are kept consistent with PSC1. Figure 4 shows that proposed scores have better accuracy and F1-score in comparison to PSC1. It also reveals that using Gaussian kernel improves the accuracy of the proposed score in the cost of a few extra parameters.

Figure 4. Three different generated PSCs against ground truth.

Conclusion

The proposed index can predict the listening difficulty of L2 learners for a given word, allows learner adaptation (by learning from learner's self-annotations), and automatically generate PSC by observing a few instances of expert annotations for a target L2 proficiency.

Acknowledgments

This research was supported by Japan's MEXT Kakenhi grant #17K02925.

References

- Bloomfield, A., Wayland, S.C., Rhoades, E., Blodgett, A., Linck, J. Ross, S. 2010. What makes listening difficult? Maryland University College Park.
- Kotani, K., Yoshimi, T. 2017. Effectiveness of Linguistic and Learner Features for Listenability Measurement Using a Decision Tree Classifier. *The Journal of Information and Systems in Education* 16(1), 7-11.
- Mirzaei, M.S., Meshgi, K., Akita, Y., Kawahara, T. 2017. Partial and synchronized captioning. *ReCALL* 29(2), 178-199.
- Yoon, S. Y., Cho, Y., Napolitano, D. 2016. Spoken text difficulty estimation using linguistic features. In *Proceedings of 11th NLP-BEA Workshop*, 267-276.