A data-driven caption for L2 listening

Maryam Sadat Mirzaei, Kourosh Meshgi
RIKEN Center for Advanced Intelligence Project (AIP), Japan

https://doi.org/10.36505/ExLing-2020/11/0042/000457

Abstract

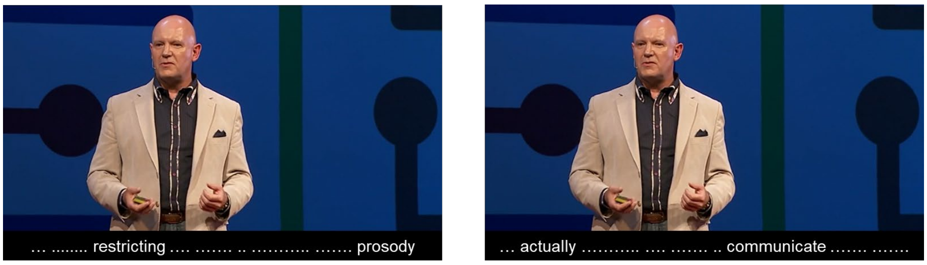
Partial and Synchronized Caption (PSC) is a tool that automatically detects difficult segments for the second language (L2) listeners and displays them in the caption while omitting easy-to-recognize cases to reduce cognitive load. Given that the number of words to be shown in this caption is limited, the main challenge lies in selecting and prioritizing difficult words. Since partialization is a classifying task, we made a dataset of labeled words in TED talks (easy vs. difficult) for a target proficiency-level. A deep classifier is trained on this dataset to automate the detection of difficult words/phrases without explicitly extracting word features. This proposed data-driven PSC outperforms its feature-based versions by adopting a selection pattern that is more similar to the annotations, capturing more complicated cases, and minimizing the false positives.

Keywords: listening difficulty, partial and synchronized caption, data-driven word selection

Introduction

The use of captions in training listening skills of L2 learners has been repeatedly studied. While full captions are considered helpful due to the simultaneous use of verbal and visual inputs according to dual coding theory (Paivio, 1990), it is criticized for overloading learner's cognitive load (Mayer & Moreno, 2003), split attention (Chang, 2009) and encouraging learners to read the text rather than listening to the audio, since it is easier for them (Leveridge & Yang, 2013).

To alleviate such problems, partial captioning has been introduced as a substitute for the full captions. The objectives of using partial captions in the literature are different. Keyword Caption (Guillory, 1998) presents only the keywords to the learner to facilitate their comprehension of the video, presenting verbal scaffold to learners regardless of their proficiency, listening difficulties, and the listening material. To address such shortcomings, we proposed Partial and Synchronized Caption (Mirzaei et al., 2018), which aims to foster L2 listening by promoting listening over reading and choosing the potentially problematic parts of listening material to appear in the caption. PSC is a type of captioning that synchronizes the words' appearance with the onset of their utterance while showing only the difficult words for the learners. This caption provides the minimal but necessary scaffold for L2 listeners to facilitate their listening, lower their reliance on reading, and promote the word boundary detection.


… …….. restricting …. ……. .. ……….. ……. prosody
… actually ……….. …. ……. .. communicate ……. …….

Original sentence: “*it's actually restricting your ability to communicate through prosody*”
How to speak so that people want to listen -Julian Treasure, TED 2013

Figure 1. Word selection for PSC: (left) shown words are labeled difficult for target learners, (right) the selected words are considered easy (and not useful) for them.

One of the challenges of making such a caption is to detect potential listening difficulties of the learners (Figure 1). Early versions of PSC use hand-crafted criteria to automatically omit easy-to-recognize words while identifying and showing difficult words to ease the cognitive load of L2 listeners. In those versions, difficult words/phrases were selected if they had a low frequency, high speech rate, or if they were considered as specific terminologies. Moreover, to add acoustically difficult words, it included special cases where automatic speech recognition errors and L2 learners’ mistakes were highly correlated (e.g., breached boundaries). However, different aspects of a word in caption may contribute to its difficulty. Thus making an exhaustive set of rules to capture such complex relation is deemed intractable and daunting. By reframing the problem to an easy-difficult word classification, the problem can be solved with state-of-the-art machine learning tools.

To this end, we prepared an annotated TED speech corpus for a target L2 proficiency level, trained a deep classifier on the speeches’ transcripts to classify its easy and difficult words, and generated PSC using this classification. This trained predictor is then used on unseen videos, and the word selection is examined against the rule-based PSC and the annotated labels (ground truth).

Method

We prepared a corpus of TED talk videos delivered by native speakers with a cumulative length of 93 minutes, including more than 10,000 words. The videos are forced aligned with Kaldi ASR system in word-level, and their show/hide labels are provided by two annotators targeting intermediate proficiency. Our annotators were L2 instructors with clear annotation guidelines and identical instructions on the criteria to label the words. The resulting annotation has κ=0.83 Cohen inter-annotation agreement.

These labels, along with the words, were used to train a classifier to predict difficult words for the target proficiency level of L2 learners. A deep text classifier is trained to classify each word into easy and difficult categories. In this classifier, words are fed to an embedding layer that maps the vocabulary of the corpus into a compact representation, increasing the classifier's accuracy. To (partially) account for acoustic aspects of the difficulties, the word embedding is augmented with the speech rate feature. This augmented encoding is then fed to two stacked LSTM layers (Hochreiter & Schmidhuber, 1997). These layers capture the context in which the word is used and provide the encoded sequence context to the following fully connected layers. The two fully connected layers employ Batch Normalization and DropOut (0.5). The processed word is finally passed to a SoftMax layer, which decides if the word should be shown (i.e., difficult word for target proficiency) or not. The model is implemented in TensorFlow.

Results and discussion

We compared our proposed data-driven PSC with the feature-based methods. In the case of feature-based PSC, the pool of features covers lexical (e.g., frequency, specificity, length, syllables), syntactic (e.g., part-of-speech, dependency parse relations), semantic (e.g., polysomic words, co-references, idiomaticity), and acoustic or perceptual complexities (e.g., speech rate, breached boundaries, negatives). Three different classifiers (SVM, Naïve Bayes, and Decision Tree) have been trained on the dataset with all these features, and the labels annotators agreed upon. Each version of PSC predicts the difficulty labels, and their performances are compared with the rule-based PSC (with speech rate, specificity, and word frequency) via 5-fold cross-validation.

Table 1. Comparing the performance of proposed data-driven PSC with feature-based PSC on the annotated dataset for intermediate learners (%)

	Easy words		Difficult words		Metrics			
Caption	Show	Hide	Show	Hide	Precision	Recall	Accuracy	Sensitivity
PSC-Rule	11.5	56.6	15.3	16.6	57.1	48.0	71.9	83.1
PSC-DT	9.2	60.9	10.3	19.6	52.8	34.4	71.2	86.9
PSC-NB	**2.4**	**71.5**	12.5	13.6	**83.9**	47.9	84.0	**96.8**
PSC-SVM	8.5	69.5	15.1	6.9	64.0	68.8	84.6	89.1
Proposed	6.7	70.8	**16.3**	**6.7**	70.9	**72.4**	**87.1**	91.4

Table 1 shows that the Naïve Bayes classifier better handles the easy words, whereas the proposed method performs better in dealing with difficult words. It also shows that these two classifiers have superior performance compared to DT, SVM, and the rule-based PSC. Since the caption should prioritize showing difficult words to L2 learners, our proposed method is more desired.

By analyzing the errors of the classifiers, several interesting findings have been observed: *(i)* Some words are considered difficult in one context but not others. Feature-based methods calculate lexical features (word frequency and specificity) regardless of the word context, which leads to some false positives in detecting difficult words, *(ii)* Since the annotators label difficult words based on the context, they consider some additional features (such as word surprisal that depends on the context) in labeling. The proposed method that uses the word context is capable of detecting such cases. *(iii)* Without adding a speech rate feature, the proposed method has difficulty in correctly identifying words that are acoustically difficult such as those uttered excessively fast. However, by adding the speech rate to the word encoding in our method, some of these cases are handled properly. On the other hand, the speech rate alone is not enough to cover a wide range of acoustically difficult words. Handling such cases requires the inclusion of more speech-driven features to the pipeline.

In summary, the proposed method is capable of predicting the listening difficulty of L2 learners for a given the word and is able to automatically generate PSC by observing a few instances of expert labels for a target L2 proficiency level. This method is extendable to allow for learner adaptation by providing annotations on different proficiency levels. It is anticipated that with more training samples, this data-driven PSC can provide a better form of assistance to foster listening to authentic materials for L2 learners at different levels.

References

Chang, A.C.S. 2009. Gains to L2 listeners from reading while listening vs. listening only in comprehending short stories. System, 37(4): 652–663.

Guillory, H.G. 1998. The effects of keyword captions to authentic French video on learner comprehension. Calico Journal, 15(1–3): 89–108.

Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), 1735-1780.

Leveridge, A.N., Yang, J.C. 2013. Testing learner reliance on caption supports in second language listening comprehension multimedia environments. ReCALL, 25(2): 199–214.

Mayer, R.E., Moreno, R. 2003. Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist, 38(1): 43–52.

Mirzaei, M.S., Meshgi, K., Kawahara, T. 2018. Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening. Computer Speech & Language, 49, 17-36, Elsevier.

Paivio, A. 1990. Mental representations: A dual coding approach. Oxford University Press.