The Significance of Fillers in Textual Representations of Speech Transcripts

Abstract

This paper investigates the role of fillers within text-based representations of speech transcripts. While often ignored in Spoken Language Understanding tasks, we demonstrate that these elements, such as "um" or "uh," when incorporated using deep contextualized embeddings, enhance the modeling of spoken language. This is further shown through improvements in downstream tasks like predicting a speaker's stance and their expressed confidence.

1 Introduction

This paper addresses the critical role of disfluencies, specifically fillers, in spoken language processing. Disfluencies, which encompass phenomena like silent pauses, word repetitions, or self-corrections, are inherent to spoken language. Fillers, a type of disfluency, often manifest as sounds like "um" or "uh," serving to bridge pauses during utterances or conversations.

While prior research has demonstrated the efficacy of contextualized embeddings pre-trained on written text for adapting to smaller spoken language corpora, these models typically exclude fillers and disfluencies in pre-processing. This practice is at odds with linguistic research, which considers fillers to be informative and integral to spoken language. Existing methods for analyzing fillers primarily rely on handcrafted features. Furthermore, pre-trained word embeddings trained on written text have shown poor performance in representing spontaneous speech words like "uh," as their meaning varies significantly in spoken contexts. In this work, we explore the use of deep contextualized word representations to model fillers. We assess their value in spoken language tasks without relying on manual feature engineering.

The core motivation of this study stems from the following observations: First, fillers are essential to spoken language. For instance, speakers may employ fillers to signal the linguistic structure of their utterances, such as difficulties in choosing vocabulary or to indicate a pause in their speech. Second, research has connected fillers and prosodic cues to a speaker's Feeling of Knowing (FOK) or expressed confidence, signifying a speaker's commitment to a statement. Fillers and prosodic cues influence a listener's perception of a speaker's expressed confidence, known as the Feeling of Another's Knowing (FOAK). Finally, fillers have been successfully applied in stance prediction, which gauges a speaker's subjective attitude.

Therefore, we intend to validate these observations by exploring how to efficiently represent fillers automatically. Our key contributions are: (1) Fillers convey useful information that can be harnessed through deep contextualized embeddings to improve spoken language modeling and should not be discarded. We also investigate the best filler representation strategies for Spoken Language Modeling (SLM) and examine the learned positional distribution of fillers. (2) In a spontaneous speech corpus of monologues, we show that fillers serve as a distinctive feature in predicting both a speaker's perceived confidence and their expressed sentiment.

2 Models and Data Description

2.1 Model Description

In this work, we focus on the two fillers "uh" and "um." To generate contextualized word embeddings for fillers, we use Bidirectional Encoder Representations from Transformers (BERT), given its state-of-the-art performance in several NLP tasks and its enhanced ability to integrate context compared to Word2Vec.

2.1.1 Spoken Language Modeling

We utilize a masked language modeling (MLM) approach for Spoken Language Modeling. This involves masking some input words at random and then attempting to predict those masked tokens. This is a standard way of pre-training and fine-tuning BERT. In our case, this method will be used to fine-tune a pre-trained BERT model on a spoken language corpus. Each experiment involves a token representation strategy i and a pre-processing strategy S_i.

The token representation strategies are essential for our goal of learning the distribution of fillers using BERT. The three token representation strategies are outlined as follows: T_1 involves no special processing for the fillers and BERT is left to use its prior understanding of fillers to model language. In T_2, "uh" and "um" are marked with specific filler tags to distinguish them from other tokens, with each filler represented as separate tokens. This strategy encourages BERT to learn new embeddings that emphasize filler context and position. In T_3, both fillers are represented as the same token, indicating that they carry the same meaning. Table 1 gives a concrete example of this process.

2.1.2 Pre-processing

We investigate the impact of three pre-processing strategies denoted by S_1, S_2 and S_3. In S_1, all fillers are removed from the sentences during both training and inference. In S_2, fillers are kept during training, but removed during inference. In S_3, fillers are preserved during both training and inference. For each combination of pre-processing and token representation strategies, we fine-tune BERT using the Masked Language Model objective like the original BERT paper. If fine-tuning is not performed the training data of S_1 and S_2 are equivalent. We evaluate the model performance in language modeling using perplexity (ppl).

2.1.3 Confidence and Sentiment Prediction

In tasks of confidence prediction and sentiment analysis, our objective is to use BERT's text representations, which include fillers, to predict a confidence/sentiment label. We add a Multi-Layer Perceptron (MLP) to BERT, which may have been fine-tuned using MLM. The MLP is trained by minimizing the mean squared error (MSE) loss. These experiments adopt the same token representation and pre-processing techniques discussed in Section 2.1.1.

2.2 Data Description

We use the Persuasive Opinion Mining (POM) dataset which contains 1000 English monologue videos. The speakers recorded themselves giving a movie review. The movies were rated between 1 (most negative) and 5 stars (most positive). The videos were annotated for high-level attributes such as confidence, where annotators rated from 1 (not confident) to 7 (very confident). Similarly, sentiment was scored by annotators between 1 (strongly negative) to 7 (strongly positive).

This dataset was chosen for several reasons: (1) The corpus contains manual transcriptions with fillers "uh" and "um," where approximately 4% of speech consists of fillers. Additionally, sentence markers are transcribed, with fillers at sentence beginnings if they occur between sentences. (2) The dataset includes monologues, where speakers are aware of an unseen listener, thus we can concentrate on fillers in speaker narratives. (3) The sentiment/stance polarity was clearly defined by choosing only reviews that were rated with 1-2 or 5 stars for annotation purposes. (4) FOAK, measured by confidence labels, has high inter-annotator agreement. More details can be found in supplementary materials. The confidence labels are the root mean square (RMS) values of labels given by 3 annotators. The sentiment labels are the average of the 3 labels.

Token. Raw	
(umm) Things that (uhh) you usually wouldn’t find funny were in this movie.	[’umm’, ’things’, ’that’, ’uh’, ’you’, ’usually

Table 1: Filler representation using different token representation strategies

3 Experiments and Analysis

3.1 Fillers Can Be Leveraged to Model Spoken Language

Language Modeling with fillers. We examine language model (LM) perplexity using various pre-processing strategies, using a fixed token representation strategy of T_1. The results in Table 2(a) compares S1, S2 and S3. By keeping fillers during both training and inference, the model reaches a lower perplexity, with a reduction of at least 10%. Therefore, fillers provide information that BERT can effectively use.

The fine-tuning procedure improves the language model’s perplexity. Additionally, even without fine-tuning, S_3 outperforms S_1 and S_2 by reducing perplexity when fillers are used. This implies that BERT has prior knowledge of spoken language and uses the fillers.

Consequently, fillers can reduce uncertainty of BERT for SLM. This is not an intuitive outcome; one might assume that removing fillers during training and inference would decrease perplexity. The fact that S_3 exceeds other preprocessing methods shows that the Masked Language Model (MLM) process effectively learns this filler information.

Best token representation: The results presented in Table 2(b) reveal that T_1 outperforms other representations when fine-tuning. Given the limited data and high BERT embedding dimensionality (768), retaining existing representations with T_1 is better than learning representations from the scratch. Interestingly, T_2 and T_3 perform similarly. The hypothesis is that the difference between "uh" and "um" lies only in the duration of the pause, which cannot be captured in text. Considering these results, T_1 is fixed as the token representation strategy in all subsequent experiments.

Learned positional distribution of fillers: We further test our model’s learning of filler placement. We fine-tune BERT using a filler to determine where the model believes the fillers most likely reside. Given a sentence S with length L, we introduce a mask token after the word j and obtain S*. We then compute the probability of a filler in position j+1.

Specifically, we calculate P([MASK=filler] | S), as depicted in Figure 1. Then, we plot the average probability of the masked word being a filler given its sentence position in Figure 2. The fine-tuned BERT model with fillers predicts a high probability of fillers occurring at the beginning of sentences. This pattern is consistent with filler distribution in the dataset. The fine-tuned BERT without fillers, predicts constant low probabilities. Given that we only know sentence boundaries we still manage to observe that the model captures a similar positional distribution of fillers that are found in other works.

(a) LM Task				(b) Best token representation		(c) FOAK and Sentiment	
Fine	Setting	Token	Ppl	Setting	Token	FOAK	Sent
3*w/o	S1	T1	22	3*S_3	T_1	1.47	1.98
	S2	T1	22		T_2	1.45	1.75
	S3	T1	20		T_3	1.30	1.44
3*w	S1	T1	5.5	3*S_3	T1	1.32	1.39
	S2	T1	5.6		T2	1.31	1.40
	S3	T1	4.6		T3	1.24	1.22

Table 2: From left to right, the (a) LM Task, (b) Best token representation, (c) MSE of Confidence (FOAK) and the Sentiment (Sent) prediction task. Highlighted results exhibit significant differences (p-value < 0.005).

1.	(umm)	I	thought	this	movie	was	really	bad
2	I	thought	=	this	movie	was	really	bad
3.	I	thought	this	movie	[MASK]	was	really	bad

Table 3: Predicting the probability of a filler, where 1. Raw input, 2. Pre-processed text with the filler removed, and 3. Illustrates the [MASK] procedure for predicting the probability of a filler at position 5

3.2 Fillers are a discriminative feature for FOAK and stance prediction

We look at the impact of fillers on two downstream tasks: FOAK prediction and sentiment analysis. Psycholinguistic studies have found a link between fillers and expressed confidence. Prior work has linked fillers and a speaker's expressed confidence in the narrow field of QA tasks. Fillers have also been used to predict stance. In this work, we present data that suggests fillers play a role in predicting a speaker's expressed confidence and their stance.

Table 2(c) shows that S_3, both with and without fine-tuning, reduces the MSE compared to S_1 and S_2. S_1 and S_2 have similar MSE since they remove fillers during inference. S_2 has a higher MSE, possibly due to the mismatch between training and test datasets. This demonstrates that fillers can be a discriminative feature in FOAK and stance prediction.

Does using fillers always improve results for spoken language tasks? In the subsection 3.1, we observe that including fillers reduces MLM perplexity. An assumption is that that downstream tasks would also benefit from the inclusion of fillers. However, we notice that when predicting speaker persuasiveness, the fillers are not a discriminative feature, following the same procedure as outlined in subsubsection 2.1.2.

4 Conclusion

This paper demonstrates that retaining fillers in transcribed spoken language when using deep contextualized representations can improve results in language modeling and downstream tasks such as FOAK and stance prediction. We also propose and compare several token representation and pre-processing strategies for integrating fillers. We plan to extend these results to consider combining textual filler-oriented representations with acoustic representations, and to further analyze filler representation learned during pre-training.