Speech Disfluencies occur at Higher Perplexities

Priyanka Sen
Amazon Alexa
sepriyan@amazon.com

Abstract

Speech disfluencies have been hypothesized to occur before words that are less predictable and therefore more cognitively demanding. In this paper, we revisit this hypothesis by using OpenAI's GPT-2 to calculate predictability of words as language model perplexity. Using the Switchboard corpus, we find that 51% of disfluencies occur at the highest, second highest, or within one token of the highest perplexity, and this distribution is not random. We also show that disfluencies precede words with significantly higher perplexity than fluent contexts. Based on our results, we offer new evidence that disfluencies are more likely to occur before less predictable words.

1 Introduction

Speech disfluencies occur naturally in spontaneous speech. Disfluencies such as filled pauses ('uh', 'um'), repetitions ('about about eight months ago'), and repairs ('about eight days I mean months ago') are estimated to occur in 6% of words in spoken English (Kasl and Mahl, 1965; Tree, 1995). In 1954, Lounsbury (1954) hypothesized a relationship between disfluencies and the likelihood of the next word. He proposed that speakers have habitual ways of speaking, and the more unexpected a word given the context, the greater the likelihood of a disfluency. Lounsbury did not test this hypothesis, saying that calculating the probability of every word in every context was, at the time, "an impossible task" (Lounsbury, 1954). Since then, several studies have found that disfluencies occur before less predictable words (Tannenbaum et al., 1965; Beattie and Butterworth, 1979; Siu and Ostendorf, 1996; Arnold et al., 2007).

In this short paper, we revisit Lounsbury (1954)'s hypothesis with newer NLP technology. Using OpenAI's GPT-2 (Radford et al., 2019), a neural language model, we calculate the predictability of words in disfluent sentences using language model perplexity. On the Switchboard corpus (Godfrey et al., 1992), a large-scale spoken language dataset, we find that 22% of disfluencies precede the word with the highest perplexity (i.e. the lowest probability), 51% of disfluencies occur either at the highest, second highest, or within one token of the highest perplexity, and this distribution is not random. We also find that words preceded by a disfluency have significantly higher perplexity than words in fluent contexts. Based on these findings, we offer new evidence of a relationship between disfluencies and less predictable words and conclude with suggested applications in NLP.

2 Related Works

Cognitive load has often been studied as a factor that affects disfluencies (Corley and Stewart, 2008). Disfluencies are found more often before longer sentences (Shriberg, 1994), in new or unfamiliar contexts (Barr, 2001; Merlo and Mansur, 2004), and when speakers are performing more challenging tasks (Oviatt, 1995). Lounsbury (1954) suggested that the likelihood of a word also affects disfluencies. Early studies evaluating disfluencies and the probability of a word used the Shannon guessing technique (Shannon, 1951) or the Cloze procedure (Taylor, 1953). In these techniques, a spoken text was transcribed and given to judges with missing words. In the Shannon guessing technique, judges guessed each word given the preceding context. In the Cloze procedure, every *n*th word was deleted, so judges had both left and

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 92–97
Barcelona, Spain (Online), December 12, 2020

right context. The probability of a word was based on how many judges could correctly guess it. With the Shannon guessing technique, Goldman-Eisler (1958) found that words preceding a silent pause had lower probability than words without a pause. Using the Cloze procedure, studies found that disfluencies occurred before words with lower probability than fluent contexts (Tannenbaum et al., 1965; Cook, 1969). Beattie and Butterworth (1979) compared word frequency and contextual probability and found that disfluencies were more likely to occur before words with lower contextual probability, even when word frequency was held constant.

With the release of larger-scale datasets, contextual probability no longer needed to be hand-annotated but could be calculated in a large enough corpus. The probability of a word given its previous word was estimated by counting the number of times both words occurred together divided by the number of times the first word occurred. This could be extended to the first *n* words in an n-gram language model. Using large corpora and n-gram language models, studies again found that disfluencies occur before words with significantly lower probability (Shriberg and Stolcke, 1996), probability varies for different disfluency types and positions (Siu and Ostendorf, 1996), and disfluencies tend to be longer before lower probability words (Harmon and Kapatsinski, 2015).

In related psychology works, eye-tracking studies have shown that listeners are primed to anticipate low frequency words when hearing a disfluency (Arnold et al., 2007; Watanabe et al., 2008) and speakers are more likely to hesitate before low frequency words (Hartsuiker and Notebaert, 2009; De Jong, 2016). Many psychology studies, however, focus more on word frequency rather than contextual probability.

3 Method

To calculate the probability of a word, we use OpenAI's GPT-2 (Radford et al., 2019), a large-scale neural language model that has achieved state-of-the-art results on various NLP tasks. GPT-2 uses a transformer-based architecture with 1.5 billion parameters and is trained on 8M documents. To our knowledge, we are the first to evaluate the relationship between disfluencies and contextual probability with a language model of this scale. To calculate probability, we use the perplexity returned by GPT-2. Perplexity is the inverse probability of a sequence normalized by the number of words. Due to the inverse, the lower the probability of a sequence, the higher the perplexity. The perplexity of a sequence of words *W* is calculated with the joint probabilities *P* of each word *w* using the formula:

$$Perplexity(W) = P(w_1, w_2..w_N)^{-1/N}$$

We use the implementation of GPT-2 available from HuggingFace (Wolf et al., 2019). Since GPT-2 is trained on written text and we experiment with spoken language, we fine-tune our GPT-2 model with the objective of predicting the next word given the previous words on 150K examples from Switchboard (Godfrey et al., 1992) for 2 epochs. This prevents our language model from predicting high perplexities for phrases that are common in spoken language but not in written language (e.g. "bye bye").

4 Dataset

For our experiemnts, we use the Switchboard corpus (Godfrey et al., 1992), which was built by asking volunteers to speak to each other on the telephone about a topic assigned by a computer operator. We use the version released by Zayats et al. (2019)[1]. We modify this dataset by including 'uh' and 'um' as disfluencies, which are included in the transcriptions but not labeled as disfluencies. From this dataset, we use 150K examples to fine-tune our GPT-2 model. We hold out 10,000 disfluent sentences from the model for our experiments. We filter this held-out disfluent set by: 1) Removing sentences that are fewer than 5 words or longer than 15 words long (excluding disfluencies), as these are often incomplete (e.g. "oh uh") or run-on sentences, 2) Removing sentences where the disfluency is the first or last word of the sentence since our experiments require left and right context to measure perplexity, and 3) Removing sentences with non-consecutive disfluencies. This is done for simplicity and because a majority (70%) of sentences contain consecutive disfluencies. Statistics about our disfluent set are shown in Table 1.

[1] https://github.com/vickyzayats/switchboard_corrected_reannotated

Count	10,000
Word Count (excl disfl)	9.5 (±3.0)
Word Count (incl disfl)	12.1 (±3.6)
Disfluency Length	2.6 (±1.8)
Disfluency Position	3.3 (±2.8)

Table 1: Statistics on the set of 10,000 disfluent utterances from Switchboard used in our experiments. Values are reported as means (± standard deviation)

Sequence	Perplexity
i'd be	80.63
i'd be very	63.12
i'd be very very	68.47
i'd be very very careful	70.90
————DISFL————	
i'd be very very careful checking	167.04
i'd be very very careful checking them	120.38
i'd be very very careful checking them out	76.44

Table 2: An example of perplexity calculated for the consecutive substrings of: "i'd be very very careful {and uh you know} checking them out".

5 Experiments

For each utterance in our disfluent set, we create a fluent version by removing the disfluencies. Using the fluent versions, we calculate the perplexity of each substring starting with the first two tokens of the sentence and adding one token at a time until the sentence is complete. An example is shown in Table 2. We expect the word following the disfluency to be the most unpredictable, so in our example, we would expect highest perplexity at the word "checking".

First, we evaluate how often disfluencies occur at the most unpredictable word. Given the list of perplexities for each sentence, we measure how often the maximum perplexity occurs at the word following the disfluency. We find that only 22% of disfluencies occur before the highest perplexity. We next calculate two more lenient measures of highest perplexity. We measure both how many disfluencies occur before the second highest perplexity and how many occur within one token of the highest perplexity. We find that 15% of disfluencies occur before the second highest perplexity, and 23% occur within one token of the highest perplexity. Taken together, 51% of disfluencies occur either before the highest perplexity, the second highest perplexity, or within one token of the highest perplexity.

The histogram in Figure 1 shows the distribution of disfluencies by rank in terms of perplexity (i.e. disfluencies at 1 occur at the highest perplexity, 2 at the second highest perplexity, etc.). This figure shows that disfluencies occur most often at the highest perplexity and trend downward for lower ranks. The histogram in Figure 2 shows the disfluency distribution by distance from the highest perplexity (i.e. disfluencies at 0 are at the highest perplexity, disfluencies at 1 are 1 token away from the highest perplexity, etc.). Here we see that disfluencies occur most often between 0 to 1 tokens from the highest perplexity and this also trends downward as distance increases. Finally, the graph in Figure 3 plots by number of words how often a disfluency occurs at the maximum perplexity compared to how often we would expect it given random chance. The error bars are calculated as a binomial proportion confidence interval based on the number of examples at that word length. For example, for all 5-word sentences,

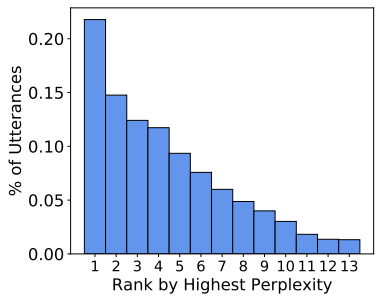

Figure 1: The distribution of disfluencies by rank in terms of highest perplexity

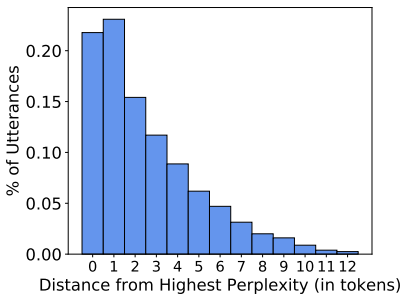

Figure 2: The distribution of disfluencies by distance from the highest perplexity

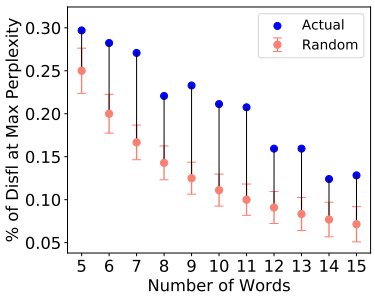

Figure 3: The occurrence of disfluencies at max perplexity compared to random chance

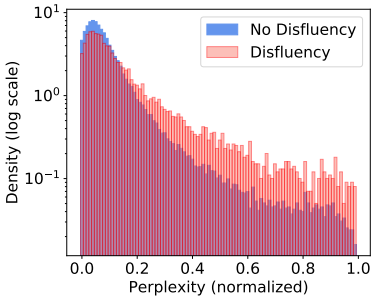

Figure 4: A histogram of the distribution of perplexities

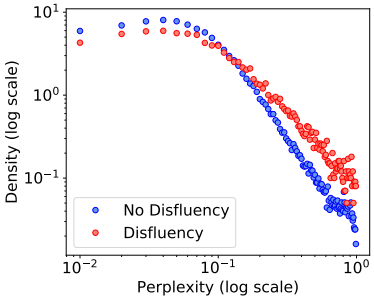

Figure 5: A log-log plot of the distribution of perplexities

	No Disfl	Disfl
Mean	0.11	0.17
Median	0.07	0.10
Q1	0.04	0.05
Q3	0.13	0.21
IQR	0.09	0.16

Table 3: Statistics on the distribution of perplexities

there are 4 possible words the disfluency could occur before (no disfluencies occur before the first word). Given random chance, we would expect a disfluency at the highest perplexity 25% of the time, but we see it closer to 30%, showing that our results are likely not a result of random chance.

Is there a bias that causes this pattern? 41% of utterances in our dataset have a disfluency at the second word, and disfluencies are known to occur more often near the beginning of the sentence (Shriberg, 1994). 46% of our highest perplexities are also at the second word. To test if we have just identified a sentence-initial bias, we run our calculations excluding examples where the disfluency or maximum perplexity is at the second word. In this filtered dataset, we find that 18% of disfluencies occur before the maximum perplexity, and 41% occur before the highest, second highest, or within one token of the highest perplexity. While these numbers are lower than on the full dataset, they still show the same pattern. This suggests that our results are not just due to a bias, and may suggest that this pattern in perplexity is related to higher planning demands at the beginning of a sentence.

Finally, we measure if disfluencies occur at higher points of statistical uncertainty compared to fluent contexts. To calculate this, we normalize the perplexities of each sentence by dividing by the sum of all perplexities in the sentence. We compare the distribution of perplexities with no disfluency against perplexities with a disfluency. The results are shown in Figures 4 and 5 and Table 3. The histogram in Figure 4 shows that the distribution of disfluent perplexities is flatter at the head and heavier in the tail, which is reflected by a higher median and third quartile in Table 3. The log-log plot in Figure 5 also shows more disfluencies at higher perplexities. After a perplexity of 0.2, disfluent and fluent perplexities diverge into two different lines, showing they follow two different distributions. The linear relationships only hold for one order of magnitude, and so they don't fully result in a power law. To calculate significance, we run both an independent t-test and a two-sided Mann-Whitney rank test, to account for the non-normal distribution. Both of these tests show significance with p-values <0.01. Based on these results, we find that disfluencies occur before words with significantly higher perplexity than fluent contexts.

6 Conclusion

In this paper, we provide new evidence using a large-scale neural language model that disfluencies occur more often before less predictable words. We consider this relationship between perplexity and disfluencies useful for applications in NLP and see the following areas as promising future directions:

- Most NLP tasks optimize to return predictions with low perplexity. However, given that disfluencies occur with higher perplexity, does preferring higher perplexity words following a disfluency give us better performance in speech or entity recognition tasks?
- Spoken language understanding tasks often need to detect disfluencies (e.g. 'a' vs. 'uh'; is a repetition part of the entity). Can we use perplexity as a signal to determine if a token is a disfluency?
- Disfluencies are uncommon, so disfluency training data is often augmented with synthetic disfluencies (Dong et al., 2019; Bach and Huang, 2019). Can we use perplexity to guide synthetic disfluency generation, and would that be more natural or useful than disfluencies inserted at random?

References

Jennifer E Arnold, Carla L Hudson Kam, and Michael K Tanenhaus. 2007. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):914.

Nguyen Bach and Fei Huang. 2019. Noisy BiLSTM-based models for disfluency detection. In *Proc. Interspeech 2019*, pages 4230–4234.

Dale J Barr. 2001. Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*, pages 597–600.

Geoffrey W Beattie and Brian L Butterworth. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, 22(3):201–211.

Mark Cook. 1969. Transition probabilities and the incidence of filled pauses. *Psychonomic Science*, 16(4):191–192.

Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Nivja H De Jong. 2016. Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, 54(2):113–132.

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6351–6358.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE Computer Society.

Frieda Goldman-Eisler. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10(2):96–106.

Zara Harmon and Vsevolod Kapatsinski. 2015. Studying the dynamics of lexical access using disfluencies. In *Papers presented at DISS 2015: The 7th Workshop on Disfluency in Spontaneous Speech*, page 41.

Robert J Hartsuiker and Lies Notebaert. 2009. Lexical access problems lead to disfluencies in speech. *Experimental Psychology*.

Stanislav V Kasl and George F Mahl. 1965. Relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, 1(5):425.

Floyd G Lounsbury. 1954. Transitional probability, linguistic structure, and systems of habit-family hierarchies. *Psycholinguistics: A survey of theory and research problems*, pages 93–101.

Sandra Merlo and Letıcia Lessa Mansur. 2004. Descriptive discourse: Topic familiarity and disfluencies. *Journal of Communication Disorders*, 37(6):489–503.

Sharon Oviatt. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–36.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

Elizabeth Shriberg and Andreas Stolcke. 1996. Word predictability after hesitations: A corpus-based study. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1868–1871. IEEE.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of California, Berkeley.

Man-hung Siu and Mari Ostendorf. 1996. Modeling disfluencies in conversational speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 1, pages 386–389. IEEE.

Percy H Tannenbaum, Frederick Williams, and Carolyn S Hillier. 1965. Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior*, 4(2):134–140.

Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Jean E Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738.

Michiko Watanabe, Keikichi Hirose, Yasuharu Den, and Nobuaki Minematsu. 2008. Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, 50(2):81–94.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Vicky Zayats, Trang Tran, Richard Wright, Courtney Mansfield, and Mari Ostendorf. 2019. Disfluencies and Human Speech Transcription Errors. In *Proc. Interspeech 2019*, pages 3088–3092.