How does BERT process disfluency?

Ye Tian, Tim Nieradzik, Sepehr Jalali & Da-shan Shiu
MediaTek Research
tiany.03@gmail.com, tim@nieradzik.me, {sepehr.jalali, DS.Shiu}@mtkresearch.com

Abstract

Natural conversations are filled with disfluencies. This study investigates if and how BERT understands disfluency with three experiments: (1) a behavioural study using a downstream task, (2) an analysis of sentence embeddings and (3) an analysis of the attention mechanism on disfluency. The behavioural study shows that without fine-tuning on disfluent data, BERT does not suffer significant performance loss when presented disfluent compared to fluent inputs (exp1). Analysis on sentence embeddings of disfluent and fluent sentence pairs reveals that the deeper the layer, the more similar their representation (exp2). This indicates that deep layers of BERT become relatively invariant to disfluency. We pinpoint attention as a potential mechanism that could explain this phenomenon (exp3). Overall, the study suggests that BERT has knowledge of disfluency structure. We emphasise the potential of using BERT to understand natural utterances *without* disfluency removal.

1 Introduction

Natural conversations are often disfluent. Consider the following utterance: "How does, I mean, *does* BERT understand disfluency?" Upon hearing this question, you understand that the speaker first tried to ask a 'how' question with a presupposition that BERT understands disfluency, but then corrected it to a yes-no question, thus removing this presupposition. Disfluent utterances like these are prevalent in natural dialogues, but rare in written texts. Recent Transformer-based language models such as BERT have amazed us in a sweep of NLP tasks requiring language understanding. Since BERT was pre-trained on written corpora, one might expect it to struggle with disfluent inputs like the one above. Traditionally, considerable effort in NLP has been devoted to disfluency detection and removal, especially in the context of dialogue systems.

But *is* disfluency removal necessary for Transformer-based language models or can they understand disfluent sentences out of the box? We approach this question from the outside in with three experiments. Experiment 1 stands outside the blackbox and explores how BERT performs behaviourally in a downstream task when presented with fluent vs disfluent language. Experiment 2 gets into the blackbox and investigates how embeddings of disfluent inputs change from the lowest to the highest layers. Finally, experiment 3 attempts to explain BERT's mechanism of disfluency processing by looking at attention on disfluent sentence parts.

We discovered that the results of all three experiments are congruent in that semantic understanding is only weakly impaired by the presence of disfluencies. Crucially, BERT represents disfluent utterances similarly to their fluent counterparts in deeper layers. This ability could be explained by the self-attention mechanism which is central to Transformed-based architectures. We hypothesise that BERT balances a trade-off between semantic selectivity and disfluency invariance[1], and that disfluency is processed similar to other *syntactic* features.

1.1 Disfluency is structured

Disfluency is ubiquitous in natural speech, found in about six out of 100 words on one estimate (Tree, 1995), and between 10% to 20% of utterances in natural dialogues on another estimate (Hough et al., 2016).

[1] Selectivity and invariance are notions more widely known in computer vision. Neurons of vertebrates develop *selectivity* to specific shapes or objects while being *invariant* to spatial and chromatic arrangements. This trade-off gives rise to object recognition robust to changes in position, rotation, occlusion and contrast. Invariance and selectivity are equally important in language. Since the essence of a sentence is found in its meaning, a robust model should develop selectivity to

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 208–217

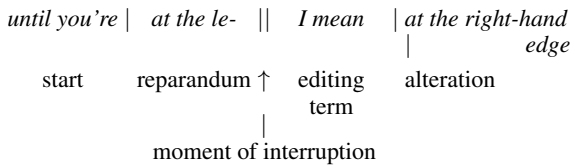

Figure 1: Structure of disfluency

Disfluencies have a consistent structure (Figure 1). They typically contain a moment of interruption, a reparandum, an editing term and an alteration (Shriberg, 1994), out of which only the moment of interruption is obligatory. Disfluencies can be *forward-* or *backward-looking* (Ginzburg et al., 2014). They are forward-looking when an utterance is interrupted by a filled or a silent pause, but are continued without an alteration. Disfluencies are backward-looking when an utterance is interrupted and replaced with an alteration that refers back to an already uttered reparandum.

This study focuses on three types of *backward-looking* disfluencies: revision, repetition and abandonment.

- A **revision** contains a reparandum and an alteration, which are both different. In the following example, "Paris" is the reparandum, "Prague" is the alteration and "I mean" is an editing term (Tian et al., 2015):
  — *"I went to Paris, I mean, Prague last week".*

- A **repetition** contains a reparandum and an alteration, and the two are the same. In this example, the first "what's your" is the reparandum and the second the alteration:
  — *"What's your, what's your old address?".*

- An **abandonment** contains only a reparandum, but no alteration. In this example, "shall we" is the abandoned reparandum, "actually" is an editing term and there is no alteration:
  — *"Shall we, actually, what's the weather like tomorrow?"*

We chose to focus on backward-looking disfluencies because they are semantically more complex than forword-looking ones. For forward-looking disfluencies, a model only needs to ignore silent or filled pauses and most commercial Automatic Speech Recognition (ASR) systems can already cope with filled pauses such as 'um' and 'uh'. For backward-looking disfluencies, there are several components such as reparanda, alterations and editing terms. Thus, a robust language model would need to not only recognise the disfluent components, but also know how they relate to each other as well as to the rest of the sentence.

1.2 Motivation

The motivation of this study is twofold: We want to explore the inner workings of BERT on disfluency processing, and we want to challenge the commonly-held belief that disfluency removal is necessary for dialogue systems.

Disfluency is rarely noise. It can aid comprehension and contribute to communicative meaning. For example, upon hearing "we believe, well, I believe that aliens exist", you understand that by changing "we believe" to "I believe", I communicate that I retract my implication of this belief being shared, to which you can respond "no, no, I believe it too". This reply would not make sense if my original utterance was the fluent counterpart "I believe that aliens exist".

Psycholinguistics studies have shown that participants anticipate more complicated concepts after a filled pause (Arnold and Tanenhaus, 2011); they remember the story better if it was told with disfluencies rather than without (Fraundorf and Watson, 2011). The processing of the reparandum helps identify the repair and has positive effects on comprehension (Shriberg, 1996). Ginzburg et al. (2014) point out that there is a continuity between self-repair and other repair types in dialogues.

Humans adapt their speech patterns to their conversational partners. Studies show that human participants tend to be more fluent when addressing a computational dialogue system than in human-human dialogues (Healey et al., 2011). However, this does not mean that humans *prefer* to speak fluently to a machine. If dialogue systems become better at understanding disfluency and are able to incrementally acknowledge and respond to disfluencies, humans will likely interact more naturally with machines. This is only possible if disfluencies are retained and gracefully handled by dialogue systems.

1.3 Related Work

The current study is related to both disfluency research and also to the study of the inner workings of BERT, often coined "BERTology". BERT (Devlin et al., 2019) is a large Transformer network

semantics while being invariant to disfluencies.

pre-trained on 3.3 billion tokens of written corpora including the BookCorpus and the English Wikipedia (Vaswani et al., 2017). Each layer contains multiple self-attention heads that compute attention weights between all pairs of tokens in the input. Attention weights can be seen as deciding how relevant every token is in relation to every other token for producing the representation on the following layer.

BERTology: In terms of syntax, Htut et al. (2019) showed that BERT's representations are hierarchical rather than linear. Jawahar et al. (2019a) found that dependency tree structures can be extracted from self-attention weights. On the other hand, studies on adversarial attacks (Ettinger, 2020) show that BERT struggles with role-based event prediction and negation. Syntactic information seems to be encoded primarily in the middle layers of BERT (Hewitt and Manning, 2019).

In terms of semantics, studies disagree in terms of where semantic information is encoded. Tenney et al. (2019) suggest that semantics is spread across the entire model. In contrast, Jawahar et al. (2019b) found "surface features in lower layers, syntactic features in middle layers and semantic features in higher layers".

Disfluency detection, removal and generation: Despite an abundance of research in probing the linguistic knowledge of written language in BERT, there is little work on probing the model on its knowledge of disfluency processing. The most related research is on disfluency detection and removal, which shifted from feature-based approaches (Hough, 2014) to more end-to-end systems (Lou and Johnson, 2020) in the past several years. Most studies use textual input, and train or fine-tune a seq2seq model using annotated disfluency data (Wang et al., 2017; Dong et al., 2019). Some studies take into account prosody (Zayats and Ostendorf, 2019). Some research stresses the importance of incremental disfluency detection (Shalyminov et al., 2018). A related emergent field is disfluency generation (Yang et al., 2020).

2 Experiments

2.1 Experiment 1: Behavioural study

Experiment 1 investigates how well BERT performs on a downstream task containing disfluent language without being exposed to disfluent data. Specifically, we used the Natural Language Inference (NLI) task (Bowman et al., 2015), where the model sees two sentences A and B, such as "A woman is singing" and "A young woman is singing". It then decides whether A entails B, contradicts B, or is neutral to B. The NLI task was chosen since it allows to quantify semantic understanding with a performance metric. By using an existing dataset and introducing disfluencies, we can observe the extent to which the accuracy degrades for different disfluency types.

Dataset: In order to compare the performance of BERT on fluent and disfluent pairs, we used data from the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015), which is a collection of 570,000 sentence pairs annotated with the labels "contradiction", "entailment" and "neutral". We took a subset of 100 sentences from the dataset and injected three types of disfluency using a combination of heuristics and manual methods[2]. Repetition was created by picking a random point of interruption in the sentence and by repeating the previous 2-4 words. Manual selection ensured that the points of interruption sounded natural. Revision and abandonment were manually created so that the disfluencies are natural and comparable between sentence A and sentence B in each pair. The final data set contains 100 fluent sentence pairs, each augmented three times for the disfluencies revision, repetition and abandonment. The introduced disfluencies do not alter the semantic meaning of these sentences. An example data point can be seen in table 1.

Sentence A	Fluent	A woman is hanging the laundry outside.
	Abandonment	A woman is hanging the laundry outside, and it was te-
	Repetition	A woman is hanging the laundry hanging the laundry outside.
	Revision	A woman is doing, I mean, hanging laundry inside.
Sentence B	Fluent	A woman is putting her clothes out to dry.
	Abandonment	A woman is putting her clothes out to dry, and it was te-
	Repetition	A woman is putting is putting her clothes out to dry.
	Revision	A woman is doing, I mean, putting her clothes out to dry.

Table 1: Example data point - Experiment 1 NLI.

2.1.1 Methods and Results

We used the medium-sized BERT model (`bert-base-cased`) which contains 12 layers,

[2]We also tried neural methods taking advantage of pre-trained language models. To generate revision, we masked between 2-4 tokens at an arbitrary position in the sentence and used BERT to "fill in the blank". The output was then concatenated with the rest of the sentence. This method often gave rise to unnatural disfluencies. Therefore, we did not use this method for data creation.

12 attention heads, and a total of 110M parameters. Using the Transformers Python library (Wolf et al., 2020), we trained a classifier by adding a softmax layer. The classifier was trained on the original SNLI data for one epoch with a batch size of 16. We then tested this model on fluent and their corresponding three disfluent sentences. The aim of experiment 1 is to assess how different disfluency types penalise the performance while using a model not trained on disfluent NLI sentences.

The results (figure 2) show that compared to the baseline accuracy of 87.5% for fluent sentences, the accuracy for *abandonment* drops slightly to 84.80% for abandonment, to 81.3% for *repetition* and to 80.4% for *revision*.

These findings suggest that without any fine-tuning on data containing disfluency, BERT already performs fairly well on the NLI task with disfluent data. With the caveat of the dataset being small and synthetic, the behaviour in experiment 1 leads to the hypothesis that BERT has an innate understanding of disfluencies. Can we find evidence for this understanding in a bigger and natural dataset? To answer this question, we carry out analyses on sentence embeddings in experiment 2.

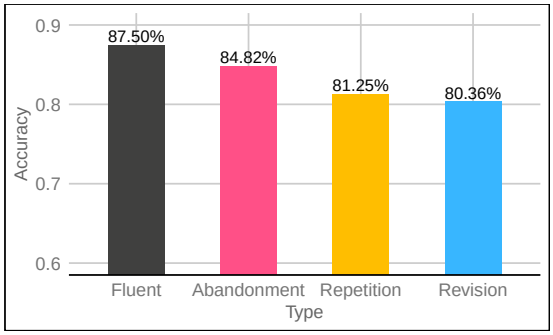

Figure 2: Experiment 1 - Model accuracy on SNLI task across Fluent, Abandonment, Repetition and Revision

2.2 Experiment 2: Inside the blackbox - Embedding Analysis

Experiment 1 shows that the performance of BERT is largely retained when the task contains a small amount of disfluency. Experiment 2 looks inside the blackbox and investigates how the embeddings of disfluent sentences change over BERT layers.

Because a disfluent sentence and its fluent counterpart are more similar in *meaning* than in *form*, we expect the sentence embeddings of the pair to be more similar in layers associated with semantic representation than layers associated with surface form and syntactic representation. If BERT indeed encodes surface form in early layers, syntax in the mid layers, and semantics in the deep layers, we should see that sentence embeddings of disfluent and fluent pairs become more similar in deep layers.

Dataset: In experiment 1, we used synthetic data. The original SNLI data is a written corpus, and disfluencies were injected manually. As such, the sentences have a different distribution from utterances appearing in natural conversations. To study the behaviour of BERT on naturally occurring disfluency, we used data from the Switchboard corpus (Godfrey et al., 1992), which is a collection of about 2,400 telephone conversations from speakers across the United States. The sentences are annotated for disfluency structure. We extracted a sample of 900 utterances balanced by disfluency type, resulting in 300 instances for *abandonment*, *repetition* and *revision* respectively. For each disfluent utterance we created a fluent counterpart by removing filled pauses, interjections and reparandam. Here is an example from this data set:

- Abandonment:
  - Disfluent: *and we just, every time you tossed the line in, you pull up a five, six, seven inch minimum bass.*
  - Fluent: *every time you tossed the line in, you pull up a five, six, seven inch minimum bass.*
- Repetition:
  - Disfluent: *um you're not supposed to, I mean, you're not supposed to eat them dead.*
  - Fluent: *you're not supposed to eat them dead.*
- Revision:
  - Disfluent: *well, today it was, I mean, the air was just so sticky, so damp.*
  - Fluent: *today the air was just so sticky, so damp.*

2.2.1 Methods and Results

Let $\mathcal{S}$ denote the dataset of all (disfluent, fluent) sentence tuples. We determine whether BERT's representation of a disfluent sentence is similar to fluent sentences using two metrics:

- Metric 1: the *raw cosine similarity* $\phi(s_d, s_f) = \frac{s_d \cdot s_f}{\max(||s_d||_2 \cdot ||s_f||_2, \epsilon)}$ computed for all $(s_d, s_f) \in \mathcal{S}$.
- Metric 2: the *cosine similarity ranking* computed for all (s_d, t_f) with $(s, t) \in \mathcal{S} \times \mathcal{S}$.

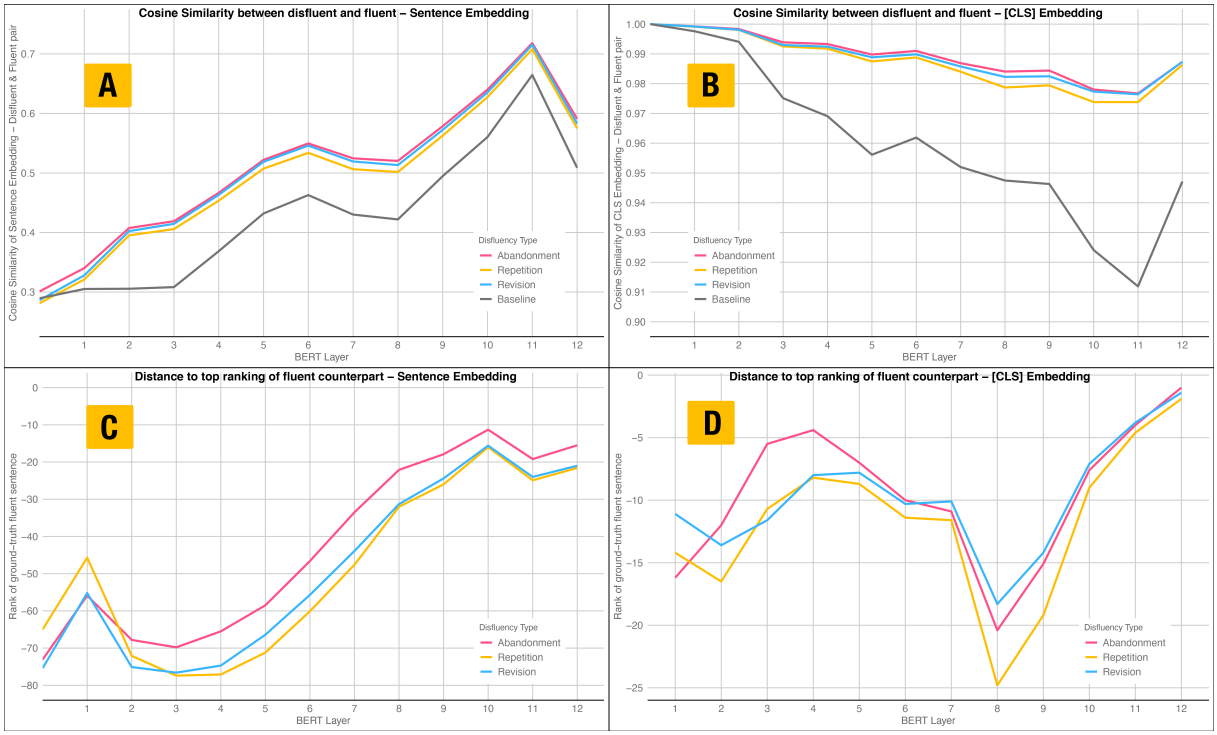

Figure 3: Experiment 2: In figures **A** and **B**, we plot the *raw cosine similarity* between each disfluent and fluent pairs, as well as between a disfluent sentence and a random fluent sentence (baseline). Figure **A** plots all sentence tokens and figure **B** plots the [CLS] token. The X axis represents layers. The Y axis represents the average cosine similarity with a range of (0,1], the closer to 1 the more similar the two vectors. In figures **C** and **D**, we plot the similarity *ranking* of the fluent counterpart - the closer to zero, the more similar the fluent counterpart compared to controls. Figure **C** ranks embeddings of all sentence tokens and figure **D** ranks the embedding of the [CLS] token. The X axis represents layers. The Y axis represents distance to top rank, so -50 means that the fluent counterpart is ranked on average 50 out of 300 in similarity.

The raw similarity (1) indicates how close a disfluent-fluent pair is in the embedding space, while a top rank in (2) determines the quality of an embedding in capturing semantic nuances. A close disfluent-fluent pair should converge to a high rank. The reasoning is that a disfluent sentence s_d is compared against all other fluent sentences t_f, some of which will be semantically similar. If the rank is high, the embeddings encode the semantic information that allows the ranking to disambiguate the correct fluent counterpart across all sentences. In other words, one could conclude that BERT's embeddings encode semantic content invariant to disfluency perturbations.

We compare two ways of sentence representation[3]: a concatenation of the embeddings of all sentence tokens, as well as the embedding of the [CLS] token. These embeddings are evaluated at all 12 layers of BERT. For comparison, we also evaluate the input vectors presented to the network.

Cosine similarity: We aggregate the activations of all sentence tokens into a single flattened vector[4]. In addition, we evaluate the activation of the [CLS] token. We calculate the cosine similarity between each disfluent sentence and its fluent counterpart. As a baseline, we calculate the cosine similarity between a disfluent sentence and a random fluent sentence. In all cases, we report the mean cosine similarity.

The results are shown in Figure 3A and 3B. Figure 3A shows that overall, the cosine similarity of a disfluent and fluent pair is higher than the baseline. The embeddings become more similar in deeper layers. An identical embedding would have a similarity of 1. At the input layer, the embeddings

[3]There is no consensus on which embeddings best represent sentence meaning. The original BERT paper (Devlin et al., 2018) proposed the hidden state of the [CLS] token on the last layer as an aggregation of sequence representation. Other studies compared pooling methods on hidden states from different layers and showed that pooling strategies are fit for downstream tasks (Ma et al., 2019).

[4]To calculate the cosine similarity between two sentences of different lengths, we pad the shorter sentence in each pair with [PAD] so that the two have the same number of tokens.

are semantically dissimilar with a mean value of 0.3. However, this value increases steadily until layer 6, plateaus on layer 7 and 8, peaks on layer 11 at around 0.72, before dropping slightly on layer 12. A similar drop was reported by Wang and Kuo (2020). The result indicates that embeddings increase in their semantic selectivity while maintaining invariance to disfluencies. We did not observe any significant difference between the three types of disfluency.

For [CLS] embedding similarity, we observe that the cosine similarity of disfluent and fluent pairs decreases as the layer gets deeper. Figure 3B shows that [CLS] embedding similarities start off at 1 on input layer, drops gradually until layer 11 to about 0.975, and increases again on layer 12. From layer 3 onwards, the [CLS] embedding similarity is higher for *abandonment* than for *repetition* and *revision*. The reason [CLS] similarity starts of at 1 is because at input layer, [CLS] embedding does not contain any information from the sentence, and is identical for all sentences. In deeper layers, [CLS] "absorbs" information and becomes more dissimilar for different sentences. Crucially, the [CLS] similarity of the baseline drops significantly over the layers compared to the three disfluent-fluent pairs.

Disfluent-fluent sentence pair ranking: In order to find out how the raw cosine similarity compares across fluent sentences for a specific disfluent sentence, we calculate the cosine similarities and compute the *rank* of the correct fluent counterpart. To reduce the computational overhead, the ranking is performed separately for each disfluency type, yielding a maximum rank of 300.

The results are shown in Figure 3C and 3D. Figure 3C shows that the similarity ranking of the fluent counterpart starts off low at around 70 on the input layer, suggesting that the tokenised surface forms of a disfluent sentence and the fluent counterpart vary significantly, which is unsurprising since disfluencies indeed render the sentences different in surface form. The ranking then sharply improves on layer 1, drops on layer 2, steadily rises all the way to layer 10, before fluctuating on layer 11 and layer 12, to a mean rank of 17 out of 300.

Why does the ranking first sharply improve on layer 1 and then drop on layers 2 to 3? We believe that this is because BERT's layer 1 primarily encodes lexical presence instead of how the tokens relate to each other. We can see that the improvement is the highest for *repetition* than for *abandonment* and *revision*. This is because in *repetition*, the tokens between the disfluent and fluent pairs are more similar. However, the advantage of *repetition* disappears from layer 2 onwards, suggesting that from layer 2, BERT starts to represent the structure and focuses less on the presence of tokens.

Among the three disfluency types, ranking for *abandonment* is the highest from layers 3 to 12. This shows that although the surface form of *abandonment* is just as different to its fluent counterpart as *revision* and *repetition*, the syntactic and semantic meaning representation of *abandonment* is more similar compared to *repetition* and *revision*, and also aligns with the results of experiment 1 (cf. figure 2).

Figure 3D shows the ranking of the [CLS] embedding of a fluent counterpart among all sentences. We removed the ranking for the input layer where the [CLS] embedding is identical for all sentences. The ranking of the [CLS] embedding of a fluent counterpart is already high at around top 15 (out of 299) on layer 1; it increases to around top 8 on layer 4, drops to top 20 on layer 8, and increases steadily until peaking on layer 12 close to the top rank.

Overall, experiment 2 shows that BERT ranks a disfluent sentence high in similarity compared to all possible fluent counterparts. In terms of the [CLS] token, the embedding on the final layer achieves top rank among 300 sentences, supporting previous studies that the final layer [CLS] embedding is a relatively good aggregation of sentence meaning. In terms of all sentence tokens, the similarity improves steadily in deeper layers, pointing towards increasing semantic selectivity and invariance to disfluencies. What could explain this selectivity-invariance tradeoff in BERT? A cornerstone of BERT is its attention mechanism which we will analyse closely in experiment 3.

2.3 Experiment 3: Attention analysis - Looking for the root cause

To understand disfluency, BERT will have to (1) identify which part in the sentence is the reparandum and which part is the alteration (if it exists), and (2) relate the reparandum and the alteration to the sentence. To investigate both aspects, we analysed attention on these disfluent segments. Previous studies show that attention weights reflect syntactic and semantic features (Clark et al., 2019). If BERT understands the structure of disfluency,

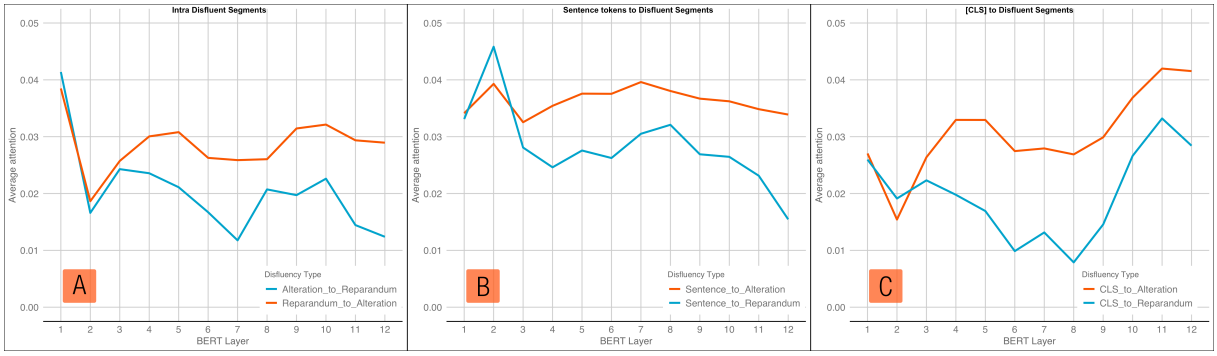

Figure 4: Experiment 3: Average Attention on each layer

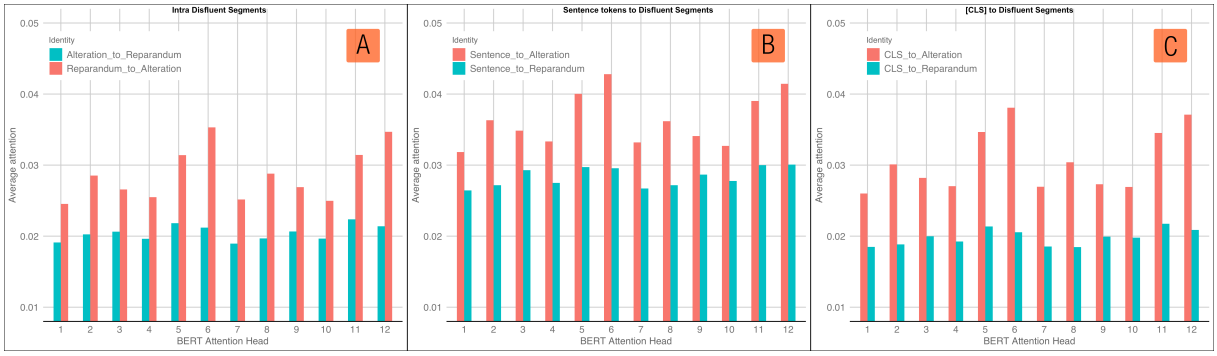

Figure 5: Experiment 3: Average Attention for each attention head

we should expect that it pays a disproportionate amount of attention to the reparandum compared to the alteration.

2.3.1 Methods and results

In order to compare the attention to reparandum and alteration, experiment 3 studies only revision and repetition. We identify the indices of the reparandum and alteration, and for each layer and each attention head, we calculated the average attention of the following:

- from the reparandum towards the alteration, and from the alteration towards the reparandum (Figure 4A, 5A)
- from all other sentence tokens towards the alteration and towards the reparandum (Figure 4B, 5B)
- from the [CLS] tokens towards the alteration and towards the reparandum (Figure 4C, 5C)

Figure 4 plots the average attention on each layer of BERT. Overall, we see that the reparandum receives less attention than the alteration from layer 3 onwards, both from all sentence tokens and from the [CLS] token. We also see that the reparandum pays more attention to the alteration than the other way around. These results suggest that in the initial layers 1-3, BERT has not distinguished the structure and different roles of the reparandum and the alteration. However, from layers 4 to 12, the reparandum contributes less to meaning representation than the alteration. The reparandum and alteration have an asymmetric relationship: the former pays attention to the later more than vice-versa.

Figure 5 plots the average attention from each attention head. Every attention head pays less attention to the reparandum than the alteration. In addition, there is more variation among attention heads on the alteration than the reparandum. Some attention heads, specifically heads 5, 6, 11 and 12 pay significantly more attention than the rest of the attention heads on the alteration. Experiment 3 once again supports the finding that the final layer [CLS] token is a good aggregation of sentence meaning. The attention heads' behaviour from [CLS] shows the same pattern as the attention from all sentence tokens.

Experiment 3 provides evidence that BERT has knowledge of the *structure* of disfluency, and this knowledge is present from the mid layers to the deep layers, akin to other syntactic and semantic knowledge. This result aligns with results from

experiments 1 and 2, and gives an insight into *how* the sentence representation of a disfluent sentence becomes more similar in deeper layers. It does so by paying less attention to the reparandum, while the reparandum attends specifically to the alteration. As a result, the meaning of the reparandum relates more weakly to the rest of the sentence compared to the alteration.

3 Discussion

Disfluencies are prevalent in natural conversations. This study investigates how Transformer-based language models such as BERT process disfluent utterances and asks whether these models have an "innate" understanding of disfluency. There are benefits of retaining instead of removing disfluencies when building dialogue systems because disfluency contributes to communicative meaning. A system that is better at understanding and responding to disfluent utterances will allow users to speak more naturally while also reducing the burden for engineers to introduce additional pipeline steps for data cleaning.

We investigated if and how BERT understands disfluency from the outside in; first by assessing the performance on a downstream task (experiment 1), then by computing sentence embedding similarities between disfluent-fluent sentence pairs (experiment 2), and finally by probing attention on disfluent segments (experiment 3).

Experiment 1 shows that without fine-tuning on disfluent data, BERT can perform fairly well on a natural language inference task containing disfluent language using a small synthetic dataset.

Experiment 2 shows that the sentence embedding of a disfluent sentence becomes more similar to its fluent counterpart the deeper the layer. Similarities of [CLS] tokens are low in earlier layers, but improve steadily in the final four layers. In addition to insights into disfluency processing, the results also suggest that layer 1 of BERT represents lexical presence without information on the relation among the tokens. The fact that pairs are most similar in the deepest layers supports previous findings that semantic meaning is more concentrated in the deeper layers of BERT.

Experiment 3 investigates why embedding similarity increases by looking at attention on disfluent segments. We found that BERT distinguishes the reparandum and alteration by paying less attention to the reparandum from layers 4 to 12.

Overall, the results are congruent in three experiments for two datasets. We conclude that BERT has knowledge of the structure of disfluency. It processes disfluency similar to other syntactic features and extracts semantic meaning by selectively attending to different parts of the disfluency at different intensities. Thus, we believe that attention is the key mechanism that modulates the selectivity-invariance tradeoff and allows BERT to embed disfluent sentences similar to fluent ones in deep layers.

4 Future work

For future studies, we could expand the scope from BERT to other Transformer language models such as DistillBERT (Sanh et al., 2019), GPT-2 (Radford et al., 2019) and XLNet (Yang et al., 2019). It would be interesting to see if language models trained with different objectives and on different data also possess the capability of resolving disfluent inputs.

In addition to more models, we could expand the scope to more languages and study if models such as multilingual BERT or MT5 (Xue et al., 2020) have knowledge of disfluency using the annotated disfluency data in German, French and Chinese from the DUEL corpus (Hough et al., 2016).

5 Conclusion

Natural conversations are filled with disfluencies such as self-repairs, repetitions and abandonment. This study shows that BERT has an out-of-the-box understanding of disfluency: it represents a disfluent sentence similar to its fluent counterpart in deeper layers. This is achieved by identifying the disfluency's structure and paying less attention to the reparandum. The results of this study raise the question whether we can use Transformer models to process disfluent utterances directly instead of first removing disfluent components in a preprocessing step. We argue that retaining disfluencies is beneficial for dialogue systems, both in terms of better capturing communicative meaning and enabling users to communicate more naturally with dialogue systems.

References

Jennifer E Arnold and Michael K Tanenhaus. 2011. Disfluency effects in comprehension: How new information can become accessible. *The processing and acquisition of reference*, pages 197–217.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Scott H Fraundorf and Duane G Watson. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language*, 65(2):161–175.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):64.

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

Patrick GT Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, pages 11–13. Citeseer.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Hough. 2014. *Modelling Incremental Self-Repair Processing in Dialogue.* Ph.D. thesis, Queen Mary University of London.

Julian Hough, Ye Tian, Laura De Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019a. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019b. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Paria Jamshid Lou and Mark Johnson. 2020. End-to-end speech recognition and disfluency removal. *arXiv preprint arXiv:2009.10298*.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal text representation from bert: an empirical study. *arXiv preprint arXiv:1910.07973*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2018. Multi-task learning for domain-general spoken disfluency detection in dialogue systems. *arXiv preprint arXiv:1810.03352*.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of international conference on spoken language processing*, volume 96, pages 11–14. Citeseer.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ye Tian, Claire Beyssade, Yannick Mathieu, and Jonathan Ginzburg. 2015. Editing phrases. *SEMDIAL 2015 goDIAL*, page 149.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Bin Wang and C-C Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

Shaolei Wang, Wanxiang Che, Yue Zhang, Meishan Zhang, and Ting Liu. 2017. Transition-based disfluency detection using lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2794.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingfeng Yang, Diyi Yang, and Zhaoran Ma. 2020. Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1450–1460.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Vicky Zayats and Mari Ostendorf. 2019. Giving attention to the unexpected: Using prosody innovations in disfluency detection. *arXiv preprint arXiv:1904.04388*.