9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

BERT Models for Spoken Learner English Disfluency Detection

Lucy Skidmore, Roger K. Moore

University of Sheffield, UK

lskidmore1@shef.ac.uk, r.k.moore@shef.ac.uk

Abstract

The automatic detection of hesitations, repetitions, and false starts commonly found in speech is a widely studied area in spoken language processing. Prior research has shown Transformer models to be highly effective at detecting such structures in native (L1) speech, however, these approaches have not yet been applied to learner (L2) data. This paper evaluates the performance of a BERT model that has been fine-tuned on spoken learner English. Results from model testing not only set a new benchmark for L2 disfluency detection comparable to that of L1 results but also show that the model can perform well on unseen corpora and across speaking activity types.

Index Terms: disfluency detection, BERT, Transformer models, learner speech, spoken language processing

1. Introduction

Collectively referred to as 'disfluencies', hesitations, repetitions and false starts are linguistic features unique to spoken language. Consider the utterance below:

$$\text{I'd like a } [\underbrace{\text{coffee}}_{\textit{reparandum}} \underbrace{\{\text{uh}\}}_{\textit{interregnum}} \underbrace{\text{cup of tea}}_{\textit{repair}}] \text{ please}$$

Disfluencies such as the example above are linguistically systematic in their structure, comprising a *reparandum* phrase, an optional *interregnum* phrase and a *repair* phrase. Examples of interregna include filled pauses such as "*uh*" seen in the example, edit terms such as "*I mean*" and finally discourse markers such as "*you know*".

The detection and subsequent removal of disfluencies from automatic speech recognition (ASR) outputs are applied to improve the performance of downstream spoken language processing tasks (see [1] as an example). In a language learning setting, disfluency detection has additionally been shown to improve the automatic detection of learner errors [2, 3] and has also been used for the automatic scoring of speaking proficiency tests [4]. However, models adapted to the L2 domain do not perform as highly as equivalent models for native speech [3, 5].

To address this issue, this paper applies a 'Bidirectional Encoder Representation from Transformers' (BERT) model [6] — the current state-of-the-art approach for L1 disfluency detection [7]— to the L2 domain. The model is fine-tuned using a dataset of spoken learner dialogues. Results from model testing not only set a new benchmark for L2 disfluency detection performance matching that of prior results on L1 data but also show that the model can perform well on unseen learner corpora and across speaking activity types.

2. Related work

Disfluency detection is a long-studied area, with various approaches having been explored in prior research. Current state-of-the-art performance is achieved using Transformer architectures, specifically BERT models which have been fine-tuned to the task of disfluency detection using conversational data. Such approaches frame the task as a sequence labelling problem, where models process whole-sentence inputs and are trained to detect only the reparandum phrase part of a disfluency.

Examples of approaches using BERT models for disfluency detection of native speech include incorporating unlabelled data with semi-supervised self-training [1], further pre-training using conversational datasets [8] and multi-task learning with joint speech recognition and disfluency detection [9]. Despite the success of the above approaches on L1 speech, with F1 scores for reparandum phrase detection reaching 92.2 [7], such approaches have yet to be explored in the L2 domain.

3. Experimentation

This experimentation evaluated the disfluency detection performance of a BERT model fine-tuned and tested using L2 data. The NICT-JLE corpus train and heldout sets established by [10] were used for model fine-tuning. Both the NICT-JLE test set and KISTEC Corpus were used for evaluation and were compared to a BERT model fine-tuned with the Switchboard Corpus [11]. The experimental details are described below and the source code and corpora used are available online[1].

3.1. BERT Model Hyperparameters

A $\text{BERT}_{\text{BASE}}$ model was used for experimentation (see [6] for an overview of the model architecture). The 'bert-base-uncased' model from the Hugging Face Transformers API[2] was fine-tuned over 3 epochs, with a train batch size of 16, AdamW-optimized weight decay starting at 0.01 and a learning rate of 5e-5.

3.2. Corpora

3.2.1. NICT-JLE Corpus

The National Institute of Information and Communications Technology Japanese Learner English (NICT-JLE) Corpus is a transcription-only corpus of 1,281 English oral proficiency tests for Japanese-speaking learners of English [12]. Conducted in an interview style between a learner and assessor, three speaking activities are carried out as part of the test: engaging in open dialogue, a role-play scenario and a picture description task.

[1] https://github.com/lucyskidmore/L2_DD_BERT
[2] https://huggingface.co/docs/transformers

3.2.2. KISTEC Corpus

The Kyoto Institute of Technology Speaking Test Corpus (KISTEC) is a transcription-only corpus of 574 English oral proficiency tests for Japanese-speaking learners of English [13]. Conducted as a computer-based test, learners are prompted to carry out three tasks: picture description, summarise a conversation that they listen to and provide an opinion on a given topic.

3.3. Data Labelling

Following [2], nested disfluencies are flattened and reparandum phrases are labelled using 'beginning-inside-outside-end-single' (BIOES) tags. See the example below:

	[I	can't	[I	I]	couldn't]	go
Tag:	B-RM	I-RM	I-RM	I-RM	E-RM	O

In the example 'RM' (reparandum) is affixed to each span (beginning, inside and end) of a reparandum phrase. Non-reparandum words are assigned 'O'. Single-word reparandum phrases are assigned 'S-RM'.

4. Results

As shown in Table 1, the BERT model fine-tuned using the NICT-JLE corpus outperforms the Switchboard baseline model for both the NICT-JLE and KISTEC test data. The F1 score of 92.8 for the NICT-JLE test set shows a significant improvement compared to prior best results using a bi-directional LSTM model [2]. Table 2 provides an overview of the model's performance according to speaking activity. As the results show, model performance is especially consistent across activities for the NICT-JLE test set. Additionally, the activities that are the least restricted (open conversation and stating an opinion) show the lowest model scores across both test sets.

Model	NICT-JLE	KISTEC
BiLSTM + Switchboard train [2]	79.8	-
BERT + Switchboard fine-tune	88.2	78.2
BERT + NICT-JLE fine-tune	**92.8**	**82.2**

Table 1: *Reparandum phrase detection F1 scores for the NICT-JLE and KISTEC corpora tested on fine-tuned BERT models compared to best scoring model from prior work.*

Activity	NICT-JLE (N=127)	KISTEC (N=574)
Conversation	91.6 (5.1)	-
Roleplay	92.2 (13.6)	-
Picture	93.5 (6.7)	76.4 (25.5)
Summary	-	77.6 (21.7)
Opinion	-	73.5 (27.9)

Table 2: *Average F1 scores and standard deviations for speaking activities in the NICT-JLE and KISTEC corpora.*

5. Conclusion

The BERT model developed and tested here provides a starting point for further experimentation on using Transformer architectures for L2 disfluency detection. Outcomes from testing have not only established a new benchmark of performance for the NICT-JLE corpus (matching that of models tested on L1 data) but also introduced the KISTEC Corpus for use in L2 disfluency detection research. In addition, the analysis of the results provided insight into how speaking activity type impacts model performance. To further improve model robustness to unseen data, it would be of interest for future research to test the impact of both pre-training and self-training using unlabelled speech corpora such as that of the ICNALE Corpus [14] as well as explore other types of Transformer model architectures.

6. References

[1] P. Jamshid Lou and M. Johnson, "Improving disfluency detection by self-training a self-attentive model," in *Proceedings of the 58th Annual Meeting of the ACL*. ACL, 2020, pp. 3754–3763.

[2] Y. Lu, M. J. F. Gales, K. M. Knill, P. Manakul, and Y. Wang, "Disfluency Detection for Spoken Learner English," in *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 74–78.

[3] Y. Lu, M. J. Gales, and Y. Wang, "Spoken Language 'Grammatical Error Correction'," in *Proceedings of Interspeech 2020*, 2020, pp. 3840–3844.

[4] R. Matsuura, S. Suzuki, M. Saeki, T. Ogawa, and Y. Matsuyama, "Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2022, pp. 1312–1320.

[5] R. Moore, A. Caines, C. Graham, and P. Buttery, "Incremental dependency parsing and disfluency detection in spoken learner English," in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 470–479.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*. ACL, Jun. 2019, pp. 4171–4186.

[7] N. Bach and F. Huang, "Noisy BiLSTM-Based Models for Disfluency Detection," in *Proceedings of Interspeech 2019*, 2019, pp. 4230–4234.

[8] J. C. Rocholl, V. Zayats, D. D. Walker, N. B. Murad, A. Schneider, and D. J. Liebling, "Disfluency Detection with Unlabeled Data and Small BERT Models," in *Proceedings of Interspeech 2021*, 2021, pp. 766–770.

[9] H. Futami, E. Tsunoo, K. Shibata, Y. Kashiwagi, T. Okuda, S. Arora, and S. Watanabe, "Streaming joint speech recognition and disfluency detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] L. Skidmore and R. Moore, "Incremental disfluency detection for spoken learner English," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Seattle, Washington: ACL, Jul. 2022, pp. 272–278.

[11] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1992, pp. 517–520.

[12] E. Izumi, K. Uchimoto, and H. Isahara, "The NICT-JLE Corpus: Exploiting the language learners' speech database for research and education," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[13] K. Kanzawa, Y. Kobayashi, J. Lee, H. Mitsunaga, M. Mori, and Y. Tanaka, "The KIT Speaking Test Corpus (KISTEC)," 2022. [Online]. Available: https://kitstcorpus.jp/

[14] S. Ishikawa, "The ICNALE Spoken Dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews," *English Teaching (The Korea Association of Teachers of English)*, vol. 74, pp. 153–177, 2019.