# Disfluency Detection for Vietnamese

**Mai Hoang Dao**[1], **Thinh Hung Truong**[2], **Dat Quoc Nguyen**[1]
[1]VinAI Research, Vietnam; [2]The University of Melbourne, Australia
{v.maidh3, v.datnq9}@vinai.io; hungthinht@student.unimelb.edu.au

## Abstract

In this paper, we present the first empirical study for Vietnamese disfluency detection. To conduct this study, we first create a disfluency detection dataset for Vietnamese, with manual annotations over two disfluency types. We then empirically perform experiments using strong baseline models, and find that: automatic Vietnamese word segmentation improves the disfluency detection performances of the baselines, and the highest performance results are obtained by fine-tuning pre-trained language models in which the monolingual model PhoBERT for Vietnamese does better than the multilingual model XLM-R.

## 1 Introduction

Humans do not always exactly predetermine what they intend to say, hence leading to interruptions in natural conversations. This phenomena is informally referred to as *disfluency* (Godfrey and Holliman, 1993; Shriberg, 1994). Disfluencies are highly ubiquitous in human conversations. With the increasing popularity of task-oriented dialogue systems, it is essential to improve the capacity of the systems in dealing with many kinds of distractor sources. Note that a vast majority of spoken language understanding (SLU) models used in the dialogue systems are trained on well-formed input text without disfluencies. However, there is a significant mismatch between the fluent training corpora and the real-world inputs of disfluent utterances/speech transcripts for those models, resulting in serious performance degradation in practical applications. Hence, disfluency detection that identifies (and then removes) disfluencies to produce fluent versions of the disfluent inputs is a crucial component of real-world SLU/dialogue systems.

Almost all benchmark datasets for the disfluency detection task, such as Switchboard (Godfrey and Holliman, 1993), CALLHOME (Canavan et al., 1997) and Child (Tran et al., 2020), are exclusively for English. Therefore, the development of disfluency detection systems has been largely limited to the English language. From a societal, linguistic, machine learning, cultural and normative, and cognitive perspective (Ruder, 2020), it is worth investigating the disfluency detection task for languages other than English, e.g. Vietnamese. In particular, it is interesting to study whether the difference in linguistic characteristics might add difficulties to developing disfluency detection systems to non-English languages, e.g. investigating the influence of Vietnamese word segmentation (Dien et al., 2001) on the Vietnamese disfluency detection task. Despite being the 17th most spoken language in the world (Eberhard et al., 2019) with about 100M speakers, to our best knowledge, there is no previous study as well as no public dataset available for disfluency detection in Vietnamese.

We fill the gap in the literature by conducting the first empirical study for Vietnamese disfluency detection. To conduct this study, we first create a dataset for Vietnamese disfluency detection through two manual phases, including: (i) adding contextual disfluencies into an existing fluent dataset of 5871 utterances (Dao et al., 2021), and (ii) annotating the added disfluencies with two different disfluency types. On our dataset, we then formulate the Vietnamese disfluency detection task as a sequence labeling problem and empirically investigate strong baselines, including BiLSTM-CNN-CRF (Ma and Hovy, 2016) and pre-trained language models XLM-R (Conneau et al., 2020) and PhoBERT (Nguyen and Nguyen, 2020). We find that: (i) automatic Vietnamese word segmentation helps improve disfluency detection performances, and (ii) the highest performance results are obtained by fine-tuning the pre-trained language models, in which the monolingual model PhoBERT outperforms the multilingual model XLM-R. We publicly release our dataset at: `https://github.com/VinAIResearch/PhoDisfluency`.

## 2 Related work

Among disfluency detection datasets with manual annotations for English (Godfrey and Holliman, 1993; Canavan et al., 1997; Tran et al., 2020; Ostendorf and Hahn, 2013; Zayats et al., 2014), the Switchboard dataset (Godfrey and Holliman, 1993) is the most commonly used benchmark for developing and evaluating disfluency detection models. The disfluency detection models generally fall into three main categories of approaches based on noisy channel, parsing and sequence tagging. Noisy channel-based disfluency detection models use tree adjoining grammar-based channel models to assign high probabilities to exact copy reparandum words (Johnson and Charniak, 2004; Johnson et al., 2004), and also use language model scores as features to a MaxEnt reranker (Zwarts and Johnson, 2011; Jamshid Lou and Johnson, 2017). Parsing-based models detect disfluencies and the syntactic structure of the sentence utterance simultaneously (Rasooli and Tetreault, 2013; Honnibal and Johnson, 2014; Yoshikawa et al., 2016; Jamshid Lou and Johnson, 2020); however, these models require large annotated training datasets that contain both disfluencies and syntactic structures. Sequence tagging approaches formulate the disfluency detection task as a sequence labeling problem to label individual words by disfluency types or simply fluent/disfluent tags (Ostendorf and Hahn, 2013; Zayats et al., 2014; Jamshid Lou et al., 2018; Bach and Huang, 2019; Rocholl et al., 2021). Among the disfluency detection approaches, the sequence tagging ones that fine-tune pre-trained language models (Devlin et al., 2019) produce the state-of-the-art performances (Bach and Huang, 2019; Rocholl et al., 2021).

## 3 Our dataset

Our approach to creating a disfluency detection dataset for Vietnamese is first to manually add contextual disfluencies as distractors into an existing fluent dataset. This first phase is inspired by Gupta et al. (2021) who present a disfluent derivative of the question answering dataset SQUAD (Rajpurkar et al., 2016). We choose PhoATIS consisting of 5871 utterance transcripts (Dao et al., 2021) as our base fluent Vietnamese dataset. After adding disfluencies to PhoATIS, we manually annotate disfluent words using disfluency types.

### 3.1 Disfluency types

A standard annotation of disfluency structure (Shriberg, 1994) includes three annotation types: the Reparandum—to annotate word or words that the speaker intends to be abandoned or corrected by the following words; the (optional) Interregnum—to annotate filled pauses, discourse cue words and the like; and the (optional) Repair—to annotate words that are used to correct the reparandum. For example, in the utterance "cho tôi biết các chuyến bay đến đà nẵng vào ngày 12 mà không ngày 14 tháng sáu" (let me know the flights to da nang on 12th uh no 14th june): "ngày 12" (12th), "mà không" (uh no) and "ngày 14" (14th) can be labeled with types Reparandum, Interregnum and Repair, respectively. Note that as pointed out in (Ostendorf and Hahn, 2013; Zayats et al., 2016), most works on automatic disfluency detection are aimed at cleaning speech transcripts to obtain fluent versions for further processing by removing disfluent Reparandum and Interregnum words. For Vietnamese, we thus annotate data using only two disfluency types Reparandum (denoted by **RM** and illustrated in red text color) and Interregnum (denoted by **IM**, in blue text color).

### 3.2 Dataset construction

**Adding contextual disfluencies:** We divide the PhoATIS's training set into 5 non-overlapping and equal subsets and preserve its validation and test sets, resulting in 7 subsets that are used for crafting disfluencies. We employ 7 annotators who are undergraduate students strong in linguistics. Here, each annotator adds disfluent words to all fluent utterances in a subset. The annotators are required to generate a disfluent version of each original fluent utterance, which: (i) is semantically equivalent to the original one; (ii) is natural in terms of human usage, grammatical errors and meaningful distractors (i.e. the added disfluent words exist in real-world circumstances); (iii) contains disfluent words that are corrected by following intent or slot value keywords in the original utterance; (iv) contains both disfluent RM- and IM-type words where possible to obtain a non-trivial dataset.

Annotators are shown example disfluencies as illustrated in Table 1. The annotators are also asked to make sure that when removing all the added words in the disfluent version, we can obtain the exact original utterance. Once the adding process is completed, the first two authors manually verify

Example 1:

mã giá vé RM: to IM: à xin lỗi tôi nhầm ý tôi là qo nghĩa là gì

what does fare code RM: to IM: uh sorry I really mean qo stand for

Example 2:

có chuyến bay nào giữa thành phố hồ chí minh và RM: hà hà nội với một điểm dừng RM: ở sân bay IM: ừm không ở đà lạt không

is there a flight between ho chi minh city and RM: ha ha noi with a stopover RM: at airport IM: uh no at da lat

Example 3:

có RM: sân bay IM: í lộn hãng hàng không nào có các chuyến bay từ điện biên phủ RM: đến quảng ninh IM: à chính xác là đến quy nhơn khởi hành trước 6 giờ 30 phút sáng không

is there any RM: airport IM: oops airline that flies from dien bien phu RM: to quang ninh IM: no actually to quy nhon departing before 6:30 am

Example 4:

tôi muốn biết thông tin về IM: ờm chuyến bay từ hạ long RM: đến IM: ờ RM: cát bà IM: ừm không tôi quên mất đến đâu nhỉ à đúng rồi đến huế bay vào buổi sáng

i'd like information on IM: uh a flight from ha long RM: to IM: uh RM: cat ba IM: uh no I forget the destination ah actually to hue a morning flight

Table 1: Disfluent utterance examples with Reparandum (RM) annotations and Interregnum (IM) annotations in our dataset. "hồ chí minh" (ho chi minh), "hà nội" (ha noi), "đà lạt" (da lat), "điện biên phủ" (dien bien phu), "quảng ninh" (quang ninh), "quy nhơn" (quy nhon), "hạ long" (ha long), "cát bà" (cat ba) and "huế" (hue) are cities in Vietnam.

| **Statistics** | **Train** | **Valid.** | **Test** | **All** |
|---|---|---|---|---|
| (1) # Utterances | 4478 | 500 | 893 | 5871 |
| (2) # Utt. w/ RM & IM | 4447 | 499 | 891 | 5837 |
| (3) # RM | 4889 | 811 | 1049 | 6749 |
| (4) # IM | 5237 | 843 | 1135 | 7215 |
| (5) Avg. Utt. length | 22.1 | 24.1 | 22.2 | 22.3 |
| (6) Avg. RM length | 2.4 | 2.3 | 2.8 | 2.4 |
| (7) Avg. IM length | 2.8 | 2.6 | 2.9 | 2.8 |

Table 2: Statistics of our dataset. (1): The number of utterances. (2): The number of utterances that contain both RM and IM annotations. (3) and (4) denote the numbers of RM and IM annotations, respectively. (5), (6) and (7) denote the average lengths (i.e. numbers of syllable tokens) of an utterance, an RM annotation and an IM annotation, respectively.

each utterance to ensure that all the requirements are met, discuss ambiguous cases and make further revisions if needed, resulting in a dataset of 5871 disfluent utterances.

**Annotation process:** Each disfluent utterance is independently annotated by the first two authors who manually annotate disfluent words using the disfluency types RM and IM. We employ Cohen's kappa coefficient score (Cohen, 1960) to measure the inter-annotator agreement between the two annotators, obtaining a substantial agreement score of 0.78. Then the third author hosts and participates in a discussion session with the first two authors to resolve annotation conflicts, resulting in a final gold dataset of 5871 disfluency-annotated utterances. Table 1 shows examples of gold annotated disfluent utterances in our dataset.

Note that when written in Vietnamese texts, the white space is used to mark word boundaries as well as to separate syllables that constitute words. Thus, the utterances in our dataset are presented at the syllable level for convenience in annotating disfluencies (e.g. the examples in Table 1). To obtain a word-level variant of the dataset, we

perform automatic Vietnamese word segmentation by using RDRSegmenter (Nguyen et al., 2018; Vu et al., 2018). For example, a 7-syllable written text "sân bay quốc tế Tân Sơn Nhất" (Tan Son Nhat international airport) is word-segmented into 3-word text "sân_bay$_{\text{airport}}$ quốc_tế$_{\text{international}}$ Tân_Sơn_Nhất$_{\text{Tan_Son_Nhat}}$". Here, automatic word segmentation outputs do not affect the span boundaries of disfluency annotations.

### 3.3 Dataset statistics

Our disfluency detection dataset for Vietnamese contains 5871 disfluency-annotated utterances, thus having a larger number of disfluent regions than Switchboard (2159), CALLHOME (1068), and Child (525). Statistic details of our dataset are reported in Table 2.

### 3.4 Discussion

Our approach that manually adds contextual disfluencies as distractors into the fluent utterances results in an artificially generated dataset. So our dataset might not correctly or fully reflect real-world scenarios where disfluencies in real-world speech might be more complex than the added contextual disfluencies in our dataset. Note that there is only one public Vietnamese speech dataset with manual transcripts used for automatic speech recognition,[1] however, the transcripts do not contain disfluencies. Thus, we could not annotate disfluencies on a real-world dataset. Our study is an attempt to imitate real-world speech and we will compare the artificially added disfluencies with the real-world disfluencies in future work.

## 4 Experiments

### 4.1 Experimental setup

Recall that the sequence labeling approaches fine-tuning pre-trained language models produce the state-of-the-art disfluency detection performances for English (Bach and Huang, 2019; Rocholl et al., 2021). Thus we formulate the Vietnamese disfluency detection task as a sequence labeling problem with the frequently used tagging scheme BIO. On our dataset, we empirically evaluate baselines that obtain competitive or state-of-the-art performances for other Vietnamese sequence labeling tasks (Nguyen and Nguyen, 2020; Dao et al., 2021; Truong et al., 2021), to investigate: (i) the influence of automatic word segmentation on Vietnamese (here, input utterances can be represented in either syllable or word level), and (ii) the effectiveness of pre-trained language models. Our baselines include BiLSTM-CNN-CRF (Ma and Hovy, 2016) and the pre-trained multilingual language model XLM-R (Conneau et al., 2020) and the pre-trained monolingual language model PhoBERT for Vietnamese (Nguyen and Nguyen, 2020). XLM-R and PhoBERT are multilingual and Vietnamese monolingual variants of the pre-trained language model RoBERTa (Liu et al., 2019). XLM-R is pre-trained on a 2.5TB multilingual dataset that contains 137GB of syllable-level Vietnamese texts, while PhoBERT is pre-trained on a 20GB word-level Vietnamese corpus.

We compute the Micro-average $F_1$ score on the validation set after each epoch, and we apply early stopping if there is no performance improvement after 5 continuous epochs. We select the model checkpoint that obtains the highest $F_1$ score over the validation set to report the final score on the test set. All our reported scores are the average over 5 runs with 5 different random seeds. See the Appendix for implementation details.

### 4.2 Main results

Table 3 presents the final $F_1$ scores (in %) obtained by the baseline models on the test set. We report the standard $F_1$ score for each different disfluency type and the Micro-average $F_1$ score for overall measurement. As the filled pauses and discourse markers belong to a closed set of words and phrases and are easier to detect (Johnson and Charniak, 2004), it is not surprising that baseline models produce about 2+% absolute higher scores for the IM type than for the RM type.

The obtained scores are categorized into two comparable settings of using the syllable-level dataset and its automatically-segmented word-level variant for training and evaluation. We find that word-level models outperform their syllable-level counterparts, thus showing the effectiveness of automatic Vietnamese word segmentation in detecting disfluent terms, e.g. BiLSTM-CNN-CRF improves from 91.54 to 92.13. We also find that fine-tuning XLM-R and PhoBERT helps produce substantially better performance scores than BiLSTM-CNN-CRF, thus confirming the effectiveness of pre-trained language models. In addition,

[1] https://institute.vinbigdata.org/en/events/vinbigdata-shares-100-hour-data-for-the-community/

| | Model | RM | IM | Mic-$F_1$ |
|---|---|---|---|---|
| **Syllable** | BiL-CRF | 88.17 | 94.67 | 91.54 |
| | XLM-$R_{base}$ | 94.61 | 97.70 | 96.21 |
| | XLM-$R_{large}$ | 95.29 | 97.75 | 96.57 |
| **Word** | BiL-CRF | 89.44 | 94.61 | 92.13 |
| | PhoBERT$_{base}$ | **95.61** | 97.28 | 96.48 |
| | PhoBERT$_{large}$ | 95.34 | **98.13** | **96.79** |

Table 3: $F_1$ score (in %) for each disluency type and Micro-average $F_1$ scores (denoted by Mic-$F_1$) on the test set. BiL-CRF denotes BiLSTM-CNN-CRF, while **Syllable** and **Word** denote scores obtained when using syllable- and word-level dataset settings, respectively.

| | Utterance length | < 20<br>44% | [20, 30)<br>44% | ⩾30<br>12% |
|---|---|---|---|---|
| **Syllable** | BiL-CRF | 92.80 | 91.44 | 88.94 |
| | XLM-$R_{base}$ | 96.52 | 96.50 | 94.74 |
| | XLM-$R_{large}$ | 96.47 | **97.23** | 95.03 |
| **Word** | BiL-CRF | 93.44 | 92.10 | 89.20 |
| | PhoBERT$_{base}$ | 96.35 | **97.23** | 94.75 |
| | PhoBERT$_{large}$ | **96.92** | 97.09 | **95.67** |

Table 4: Mic-$F_1$ scores (in %) w.r.t. utterance lengths (i.e. the numbers of syllable tokens). The numbers (44%, 44% and 12%) right below length buckets denote the percentages of utterances belonging to the buckets.

PhoBERT does better than XLM-R ("base" versions: 96.48 vs. 96.21; "large" versions: 96.79 vs. 96.57), however, the score differences between PhoBERT and XLM-R are not substantial. It is probably because our utterances are domain-specific and contain disfluencies, while PhoBERT is pre-trained on domain-general and fluent data.

We also present the Micro-average $F_1$ scores (in %) w.r.t. utterance length buckets on the test set in Table 4. Those obtained scores generally show that the baseline models perform better when the input utterances are shorter than 30 tokens. The longer the input utterances are (i.e. longer than 30 tokens), the more ambiguous their meanings are and the more confused the baselines get.

### 4.3 Error analysis

To understand the source of error, we conduct an error analysis using the best performing model PhoBERT$_{large}$ that returns a total of 45 incorrect predictions on the validation set (average over the 5 different runs).

The first error group consists of 27/45 instances with inexact disfluency boundaries (i.e. inexact spans) overlapped with gold spans but having correct disfluency labels, while the second error group consists of 4/45 instances with the overlapped inexact spans and incorrect labels. These 27 + 4 = 31 errors are largely caused by the dropping of a reparandum-related term inside the fluent correction part, without affecting the utterance's semantic meaning, however, resulting in contextual ambiguity to the model. For example, in the utterance "tôi muốn biết giá vé hạng thương gia à nhầm phổ thông" (I would like to know the ticket price for the business class oops economy),[2] the whole phrase "hạng thương gia" (business class) is wrongly predicted as a RM while it must only be "thương gia" (business). Here, it is worth noting that the contextual ambiguity is resulted by a dropping of a possibly additional secondary term "hạng" (class) to be coupled "phổ thông" (economy), i.e. "hạng phổ thông" (economy class).

The third group of 2/45 errors with exact spans and incorrect disfluency labels does not provide us with any useful insight. The model also produces the fourth group of 9 errors where gold-annotated disfluent words/phrases are predicted with the label O. The majority of these 9/45 errors are caused by the fact that disfluencies can exist anywhere in a Vietnamese utterance, e.g. IM disfluent words can appear at the end of the utterance. For example, with the utterance "chuyến bay buổi sáng à không tôi đang vội chuyến bay đầu tiên nhé" (morning flight uh no I'm in hurry first flight please), the model could not predict the word "nhé" as an IM. The last error group consists of 3/45 instances where predicted disfluencies are associated with the gold label O. They are general terms such as "sân bay" (airport), "thành phố" (city) and the like, that frequently used in disfluent phrases. Thus, when occurred in the fluent parts of an utterance, these terms are likely predicted as disfluencies, leading to incorrect predictions.

## 5 Conclusion

In this paper, we have presented the first study for Vietnamese disfluency detection. We create a Vietnamese disfluency detection and empirically conduct experiments on this dataset to compare strong baseline models as well as perform detailed error analysis. Experimental results show that the input representations and the pre-trained language models have positive influences on this Vietnamese disfluency detection task.

[2]Word segmentation is not shown for simplification. Here, we also color the gold annotations.

## References


Nguyen Bach and Fei Huang. 2019. Noisy BiLSTM-Based Models for Disfluency Detection. In *Proceedings of INTERSPEECH*, pages 4230–4234.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42. Linguistic Data Consortium.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*, pages 8440–8451.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of INTERSPEECH*, pages 4698–4702.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, pages 4171–4186.

Dinh Dien, Hoang Kiem, and Nguyen Quang Toan. 2001. Vietnamese Word Segmentation. In *Proceedings of NLPRS*, pages 749–756.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2019. *Ethnologue: Languages of the World, 22nd edition*. SIL International, United States.

John Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. *Linguistic Data Consortium*.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering. In *Findings of ACL*, pages 3309–3319.

Matthew Honnibal and Mark Johnson. 2014. Joint Incremental Disfluency Detection and Dependency Parsing. *Transactions of ACL*, 2:131–142.

Paria Jamshid Lou, Peter Anderson, and Mark Johnson. 2018. Disfluency Detection using Auto-Correlational Neural Networks. In *Proceedings of EMNLP*, pages 4610—-4619.

Paria Jamshid Lou and Mark Johnson. 2017. Disfluency Detection using a Noisy Channel Model and a Deep Neural Language Model. In *Proceedings of ACL*, pages 547–553.

Paria Jamshid Lou and Mark Johnson. 2020. Improving Disfluency Detection by Self-Training a Self-Attentive Model. In *Proceedings of ACL*, pages 3754–3763.

Mark Johnson and Eugene Charniak. 2004. A TAG-based noisy-channel model of speech repairs. In *Proceedings of ACL*, pages 33–39.

Mark Johnson, Eugene Charniak, and Matthew Lease. 2004. An improved model for recognizing disfluencies in conversational speech. In *In Proceedings of Rich Transcription Workshop*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint*, arXiv:1412.6980.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pages 1064–1074.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of EMNLP*, pages 4079–4085.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of EMNLP*, pages 1037–1042.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of LREC*, pages 2582–2587.

Mari Ostendorf and Sangyun Hahn. 2013. A sequential repetition model for improved disfluency detection. In *Proceedings of INTERSPEECH*, pages 2624–2628.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP*, pages 2383–2392.

Mohammad Sadegh Rasooli and Joel Tetreault. 2013. Joint Parsing and Disfluency Detection in Linear Time. In *Proceedings of EMNLP*, pages 124–129.

Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency Detection with Unlabeled Data and Small BERT Models. In *Proceedings of INTERSPEECH*, pages 766–770.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. https://ruder.io/nlp-beyond-english/.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California.

Trang Tran, Morgan Tinkler, Gary Yeung, Abeer Alwan, and Mari Ostendorf. 2020. Analysis of Disfluency in Children's Speech. In *Proceedings of INTERSPEECH*, pages 4278–4282.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of NAACL*, pages 2146–2153.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of NAACL: Demonstrations*, pages 56–60.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Joint Transition-based Dependency Parsing and Disfluency Detection for Automatic Speech Recognition Texts. In *Proceedings of EMNLP*, pages 1036–1041.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency Detection Using a Bidirectional LSTM. In *Proceedings of INTERSPEECH*, pages 2523–2527.

Victoria Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2014. Multi-domain disfluency and repair detection. In *Proceedings of INTERSPEECH*, pages 2907–2911.

Simon Zwarts and Mark Johnson. 2011. The impact of language models and loss functions on repair disfluency detection. In *Proceedings of ACL*, pages 703–711.

# A Appendix

### Experimental models

- BiLSTM-CNN-CRF (Ma and Hovy, 2016) represents each input token by concatenating its corresponding pre-trained token embedding and CNN-based character-level token embedding; then concatenated representations of input tokens are fed into a BiLSTM encoder to extract latent feature vectors for the input tokens; each latent feature vector is then linearly transformed before being fed into a linear-chain CRF layer (Lafferty et al., 2001) for disfluency label prediction.

- Fine-tuning XLM-R (Conneau et al., 2020) or PhoBERT (Nguyen and Nguyen, 2020) for disfluency detection is done in a common approach that uses a linear prediction layer on top of its architecture. In other words, we feed the XLM-R- or PhoBERT-based contextualized token embeddings as input for the linear prediction layer, to predict the disfluency label for each token.

| **Hyper-parameter** | **Value** |
|---|---|
| Optimizer | Adam |
| Learning rate | 0.001 |
| Mini-batch size | 36 |
| LSTM hidden state size | 200 |
| Number of BiLSTM layers | 2 |
| Dropout | [0.25, 0.25] |
| Character embedding size | 50 |
| Filter length, i.e. window size | 3 |
| Number of filters | 30 |
| W2V embedding dimension | 300 |

Table 5: Hyper-parameters for BiLSTM-CNN-CRF.

For training the baseline BiLSTM-CNN-CRF, we employ the pre-trained 300-dimensional Word2Vec syllable and word embeddings for Vietnamese from (Nguyen et al., 2020). We fix these embeddings during training. Optimal hyper-parameters that we select via performing a grid search for BiLSTM-CNN-CRF are presented in Table 5. We fine-tune XLM-R and PhoBERT for the syllable- and word-level settings, respectively, using the optimizer Adam (Kingma and Ba, 2014) with a fixed learning rate of 5e-5 and a batch size of 32 (Liu et al., 2019). Note that BiLSTM-CNN-CRF is trained for 50 epochs while XLM-R and PhoBERT are fine-tuned for 30 training epochs.