INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Disfluency Detection using a Bidirectional LSTM

Vicky Zayats, Mari Ostendorf and Hannaneh Hajishirzi

Electrical Engineering Department
University of Washington
{vzayats,ostendor,hannaneh}@uw.edu

Abstract

We introduce a new approach for disfluency detection using a Bidirectional Long-Short Term Memory neural network (BLSTM). In addition to the word sequence, the model takes as input pattern match features that were developed to reduce sensitivity to vocabulary size in training, which lead to improved performance over the word sequence alone. The BLSTM takes advantage of explicit repair states in addition to the standard reparandum states. The final output leverages integer linear programming to incorporate constraints of disfluency structure. In experiments on the Switchboard corpus, the model achieves state-of-the-art performance for both the standard disfluency detection task and the correction detection task. Analysis shows that the model has better detection of non-repetition disfluencies, which tend to be much harder to detect.

1. Introduction

A characteristic of spontaneous speech that makes it different from written text – including informal text – is the presence of disfluencies. Disfluencies include filled pauses, repetitions, repairs and false starts. Disfluencies are frequent in all forms of spontaneous speech, whether casual discussions or formal arguments [1]. They present significant challenges for some natural language processing (NLP) tasks on spoken transcripts, such as parsing and machine translation [2, 3, 4]. On the other hand, disfluencies also reflect speaker interaction [5]. Disfluency detection is most often used as a preprocessing step for NLP, where the goal is removal of the non-fluent word sequences. For extracting information about the interaction, the detection of both disfluent and correction parts can be important.

A standard annotation of disfluency structure [6] indicates the reparandum (word or words that the speaker intends to be replaced or ignored), the interruption point (+) marking the end of the reparandum, the associated repair, and an optional interregnum after the interruption point (filled pauses, discourse cue words, etc.)

$$[\text{reparandum} + \{\text{interregnum}\}\ \text{repair}]$$

Ignoring the interregnum, disfluencies can be categorized into three types: restarts, repetitions, and corrections, based on whether the repair is empty, the same as the reparandum or different, respectively. Table 1 gives a few examples. In this work, we use a slightly modified representation from [7] that distinguishes repetitions (marked by 'S') and flattens the nested structure in a sequence of repetitions, which has led to improved disfluency detection in prior work [7, 1].

Most work on automatic disfluency detection is aimed at cleaning transcripts for further processing, where only reparandum detection is of interest. In this study we are interested in both the reparandum and repair, motivated by a long term goal of understanding variability in disfluency production related to cognitive load and social context. We introduce a new approach to disfluency detection given text transcripts that leverages a Bidirectional Long-Short Term Memory (BLSTM) neural network and integer linear programming. The model achieves state-of-the-art performance on the standard Switchboard task given speech transcripts, and analyses show contributions from including pattern match features in the input.

Type	Annotation
repair	[I just + I] enjoy working
repair	[we + you'd] have to just
repair	[we want + {well} in our area we want] to
repetition	[S it's + {uh} it's] almost like
repetition	[S the + th- + the] decision was
restart	[by +] it was attached to
restart	[we would like +] let's go to the

Table 1: Examples of different types of disfluencies.

2. Related Work

Approaches to automatic disfluency detection generally fall into two categories: sequence tagging and parsing-based models. Many studies have used a sequence tagging model with begin-inside-outside (BIO) style states that label words as being inside or outside of a reparandum word sequence. The most successful approaches have leveraged discriminative models, including conditional random fields (CRFs) as a classifier [8, 9, 7, 1, 10]. In [9], Integer Linear Programming is integrated with the CRF for optimizing over the prediction sequences. An alternative for improving the CRF uses an F-score matching objective, multi-step learner and Max-Margin Markov Networks (M^3N) [11]; the objective change had the highest impact. The current best performing system uses a Semi-Markov CRF [12].

Another set of approaches leverage parsing and represent a noisy channel relationship between the reparandum and the repair [13, 14]. The noisy channel parsing models could be used for identifying repairs, though results on that task have not been reported. Incremental dependency parsing combined with disfluency removal has also been explored [15, 3]. Because incremental models do not benefit from reparandum/repair similarity cues, they tend to have lower performance than delayed decision models. Depending on the downstream application, an advantage of parsing models in general is that they jointly optimize for both parsing and disfluency detection performance. A disadvantage is that they require treebank annotation for training. Since we are ultimately interested in applying disfluency detection to a broad set of domains, we will leverage a sequence tagging approach, but we extend the label state space to separately

http://dx.doi.org/10.21437/Interspeech.2016-1247

model repetitions and repairs, as in [1].

Two recent studies have applied recurrent neural networks (RNNs) to disfluency detection. One approach explores incremental detection [16], with an objective that combines detection performance with minimal latency. Because of the latency constraints, this approach has weak performance in comparison to other studies on disfluency detection. Word embeddings learned by an RNN have also been used as features in a CRF classifier [10]. In our current study, we also use an RNN, particularly the Long-Short Term Memory (LSTM) framework, but in the standard disfluency detection paradigm (non-incremental), which allows us to use a bidirectional architecture and leverage the relatedness of repair and reparandum for repetition and correction disfluencies. Unlike [10], the RNN is the classifier, so our feature embeddings are trained in an end-to-end manner, and they also leverage pattern matching features.

While most studies of disfluency detection focus on using only text transcripts as input, it is well known that prosodic cues are useful in combination with lexical cues [17, 18, 8, 12]. Prosody can carry information that is not represented in transcripts (e.g. length of pauses, fundamental frequency trends), which is relevant for detecting interruption points. However, most studies find that the gain from combining prosodic features with lexical features is relatively small, so our current study focuses on lexical features alone. Adding prosodic information to the existing features is an easy modification with the neural network framework, which we hope to explore in future work.

3. General Framework

The standard disfluency detection task involving reparandum detection is often called "edit detection." The typical sequence tagging model represents 5 states: beginning of the edit region *BE*, inside edit *IE*, the word before the interruption point *IP*, one word edit *BE_IP* and outside of the edit (including both repairs and fluent regions) *O*. For evaluation of edit detection, all words with labels other than *O* are considered edit words. We also consider two extensions of the state space. The first extension (called explicit repair modeling) includes 8 states, adding: *C* for the repair word, and *C_IE, C_IP* for words in nested disfluencies that belong to both a reparandum and a repair. For the edit detection task, the *C_IE, C_IP* states are considered part of an edit region. Note that having explicit repair states does not allow correction detection, as defined in [1], since the repairs associated with repetitions vs. corrections are not distinguished. The expanded state space uses the extent of the correction to improve edit detection. The second extension includes 17 states for joint reparandum and correction detection, expanding all non-*O* states to separately represent repetition and non-repetition disfluencies, as in [1]. With this model, we can detect corrections and take advantage of the fact that repetitions tend to benefit from different features than other disfluencies.

As reviewed in section 2, CRFs have been used widely for disfluency detection and therefore represent a strong baseline for comparison to the new models developed here. In our work we use the CRF++ toolkit [19]. Starting from a core feature set of lexical, distance-based pattern match features and disfluency language model features used in [1] (listed in Table 2), which were engineered for multi-domain disfluency detection. In particular, features 7-17 indicate a pattern match in terms of specific words or POS tags, providing domain-independent indicators of repetitions and simple corrections. The CRF features are generated by applying feature combination functions provided by CRF++ templates to create new features within a relative time frame. For example, using the core feature 'word index,' we can construct n-gram features by applying feature functions across a local time window. A total of 258 features are generated, including combinations of different core features as well as n-grams and POS n-grams.

Core Features
1. word index
2. part of speech (POS) tag
3. is the word a filled pause
4. is the word a discourse marker
5. is the word a part of an edit word
6. is the word incomplete
7. distance to the repeated word in the following window
8. distance to the repeated bi-gram in the following window
9. distance to the repeated word in the preceding window
10. distance to the repeated bi-gram in the preceding window
11. is the POS bi-gram repeated in the following window
12. is the POS bi-gram repeated in the preceding window
13. is the word and the POS of the next word repeated within the following window
14. is the POS and the next word repeated within the following window
15. is the word bigram repeated within the next N words allowing some words to come between the two words
16. is POS trigram repeated within the N words
17. distance to the next used conjunction word
18-20. 3 language model features described in [1]

Table 2: Core features used to generate CRF features and feature embeddings.

4. Proposed Method

4.1. RNN Architectures

We use LSTM RNNs for the task of disfluency detection, since LSTMs achieve good performance in a variety of NLP sequence modeling tasks [20, 21, 22]. A typical memory cell includes gates to weigh input and history impact at a particular time, allowing the model to determine their relative importance and alleviating the vanishing gradient problem [23]. As a result, LSTMs can effectively represent longer phrases, which is useful for the disfluency detection task. For disfluency state sequence tagging, we use a softmax layer at the top layer of the LSTM.

An LSTM is a directional model and predicts a state given its previous states. For disfluency detection based on text alone, it is difficult to predict a word as disfluent by only observing the words prior to the occurrence of the interruption point. Unexpected word sequences following the interruption point and similarity between the repair and the reparandum are important indicators. Therefore, the LSTMs used here take the input sentence in reverse order. In addition, we explore use of a bidirectional LSTM (BLSTM) [24]. As shown in Figure 1, the BLSTM uses past and future states in predicting the disfluency tag of a given word. This is particularly useful for predicting both repairs and corrections (8-state and 17-state models).

4.2. Feature Embeddings

As shown in Figure 1, the input vector consists of three main components: word index, POS tag, and disfluency-based features, as listed in Table 2. The disfluency features provide useful

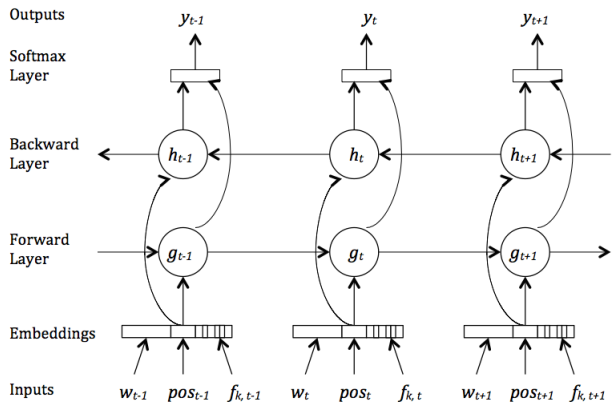

Figure 1: Sample Bidirectional RNN architecture which includes feature embeddings. w_t is a word at time t, pos_t is the POS at time t, $f_{k,t}$ is the k-th core feature at time t , and y_t is an output state (e.g. *BE_IP*).

information about word identity (filler or incomplete words) and patterns (if the exact word appeared previously in a fixed length window). We separately map one-hot representations of these features to embeddings for a dense representation, and then concatenate them to use as the input to LSTM cells. For initialization of the word embeddings, we train a backward-LSTM language model on the Switchboard corpus with disfluencies removed. The initialization for POS tag embeddings is similar, with the training text mapped to POS tags. All other parameters have random initialization. During the training of the whole neural network, embeddings are updated through back propagation similar to all the other parameters.

4.3. ILP post-processing

While the hidden states of LSTM and BLSTM are connected through time, the outputs from the softmax layer are not. This often leads to inconsistencies between neighboring labels, sometimes resulting in label sequences that are not valid paths in the state space (i.e. 'illegal'). For example, the model can output the sequence of labels *O IE IE IP*, which is not a valid sequence since disfluencies always start with *BE* tag. As such, an additional smoothing over LSTM predictions is needed. Some of the possible approaches include LSTM-CRF [25], Markov model [16] or Integer Linear Programming (ILP) [9]. In this work we use the ILP solution previously presented in [9] for disfluency detection. Since [9] uses constraints for 5 states only, we collapse the softmax proportions from the larger state space to 5 states, as in

$$\begin{aligned} P(y_t = O) &= P(y_t = O) + P(y_t = C) \\ P(y_t = IE) &= P(y_t = IE) + P(y_t = C\_IE) \\ P(y_t = IP) &= P(y_t = IP) + P(y_t = C\_IP) \end{aligned} \quad (1)$$

for the 8-state model.

5. Experiments

We assess the proposed BLSTM model for disfluency detection in experiments on the Switchboard corpus of conversational speech [26], using the standard division of the disfluency-annotated subset into training, development and test sets. The flattened version of repetition annotation provided in [7] is used.[1] As in other studies on disfluency detection, performance is measured using precision/recall of words in edit regions. In addition, we use the same measure on finer grain labels, including different types of edit regions (repetition vs. non-repetition disfluencies) and corrections.

All code is written in theano [27]. LSTM and BLSTM parameter optimization is done using Adadelta [28] with a mini-batch size of 50. We use the Switchboard development set to tune the LSTM parameters (number of dimensions) and to find an optimal stopping point for LSTM training. The dimensions of the word embeddings and the hidden dimension are separately tuned for each variation. The best number of dimensions for the BLSTM with 17 states is 100, and for all other models it is 150. The POS embedding dimension is chosen to be 5 for all models. As mentioned previously, word and POS embeddings are initialized using a backwards LSTM language model trained on cleaned-up Switchboard text, and other model parameters have random initialization. All models are trained using only sentences that have 50 or fewer words due to the high computational complexity of longer histories in the LSTM. Fewer than 1% of Switchboard sentences have length greater than 50 words. In testing, all sentences are processed. For the results described below, significance is assessed using a paired t-test on the number of errors in each sentence.

Model	P	R	F
CRF *5 states*	91.7	78.1	84.3
CRF *8 states*	91.3	77.6	83.9
CRF *17 states*	92.9	76.1	83.7
BLSTM *5 states*	93.6	79.0	85.7
BLSTM *8 states*	91.5	81.5	**86.2**
BLSTM *17 states*	90.7	81.5	85.8
BLSTM *8 states* + ILP	92.7	81.9	**87.0**

Table 3: Performance of the LSTM and Bidirectional LSTM on the dev set in the edit detection tasks.

5.1. Model Performance

Table 3 shows the CRF and BLSTM performance on the development set for the edit detection task using all three state space alternatives presented in the Section 3. As shown in the table, the BLSTM has significantly better performance in edit detection compared to the CRF ($p < 0.01$ for all cases). Moreover, the BLSTM with explicit repair states achieves the best result in the edit detection task, which we hypothesize is related to the success of the noisy channel model approach: explicitly representing the extent of the repair allows the model to match the repair to the reparandum for improved detection. Table 3 also gives the result of the ILP post-processing on the best model. Although some corrections to illegal sequences do not impact edit detection performance, many do. ILP improves the precision of our BLSTM predictions without hurting the recall, but only 0.4% of labels change so the decrease in errors is not statistically significant.

The 17-state models give slightly worse performance for edit detection, but they enable correction detection, the results of which are shown in Table 4. While the BLSTM gives significantly better overall F-score, the two models have very differ-

[1] Annotated data is available at `http://ssli.ee.washington.edu/tial/data/disfluencies/`.

Model	P	R	F
CRF *17 states*	73.2	37.5	49.6
BLSTM *17 states*	57.3	51.3	54.2

Table 4: Correction detection on the dev set using 17 states.

Model	P	R	F
Qian et al. [11]	-	-	84.1
Honnibal et al. [3]	-	-	84.1
Ferguson et al. [12] (lexical)	90.1	80.0	84.8
BLSTM 8 states	91.4	80.3	85.5
BLSTM 8 states + ILP	91.8	80.6	**85.9**

Table 5: Comparison of the BLSTM model to state-of-the-art methods in the literature on the test set.

ent precision-recall tradeoffs. Augmenting the 17-state BLSTM with ILP post-processing could potentially recover some of the precision lost in moving to the BLSTM.

5.2. Method Comparison

We evaluate our best models on the test set and compare them to recent methods in the literature leveraging only transcripts. For edit detection, we use the explicit repair state space (8 states), which achieves the best results on the development set, including results both with and without ILP post-processing. The results are shown in Table 5. Both systems beat the best prior result with lexical cues only, achieving state-of-the-art performance of 85.9. Again, the decrease in errors for ILP is not statistically significant. The BLSTMs also beat the higher performing version in [12] that leverages prosodic features (F=85.4). Incorporating prosodic features in a neural network framework is straightforward and will likely lead to an additional gain.

The 17-state BLSTM model also leads to a significant performance gain in correction detection on the test set, achieving an F-score of 57.7 compared to 49.6 for the CRF,[2] corresponding to a 16% improvement. The 17-state BLSTM finds disfluencies that the CRF misses entirely, as in the examples:

Ref: *a [drug policy there + drug testing policy] where they*
BLSTM: *a [drug policy + there drug testing policy] where they*

Ref: *so [we're + uh our discussion's] about uh the care of*
BLSTM: *so [we're +] uh our discussion's about uh the care of*

The second case could arguably be analyzed as a restart.

5.3. Ablation study

We conducted an ablation study on the effect of engineered features with the different models. Results are shown in Table 6 for edit detection (5 states) on the development set. Differences between the LSTM and CRF systems in both configurations are significant ($p < 0.05$), but the BLSTM improves over the LSTM only with the expanded feature set ($p < 0.05$). For the *words-only* cases, the CRF word features include 1-3 grams within a window of 8 around the word, whereas the LSTM and BLSTM use only the current word index and incorporate longer context through the recurrent structure. When we add POS and pattern-match features, all systems improve, but the CRF benefits much more than the other models. The impact of expanding the feature set is much greater than the different model configurations in all cases.

[2]The CRF result is the same as that reported in [1]. That work describes it as a 16-state model, since the outside-disfluency state was not counted, but it is the same as the 17-state model described here.

Model (*input*)	P	R	F
CRF (*words*)	94.4	52.8	67.7
CRF (*words+ pos + feat*)	91.7	78.1	84.3
LSTM (*words*)	87.6	71.4	78.7
LSTM (*words + pos+ feat*)	92.4	79.0	85.2
BLSTM (*words*)	87.8	71.1	78.6
BLSTM (*words + pos+ feat*)	93.6	79.0	85.7

Table 6: Comparison of 5-state CRF, LSTM and BLSTM edit detection models with different feature sets on the dev set.

Model (*input*)	Repetitions	Other	Either
CRF [1] *17 states*	94.9	61.1	83.7
BLSTM *17 states*	94.1	66.7	85.8

Table 7: F scores of different types of edits for the CRF and BLSTM on the dev set.

5.4. Repetitions vs. Non-repetitions

Repetition disfluencies are much easier to detect than other disfluencies, although not trivial since some repetitions can be fluent. In order to better understand model performance, we evaluate the 17-state models in terms of their ability to detect repetition vs. non-repetition (other) reparanda. The results are shown in Table 7, showing that the BLSTM is much better in predicting non-repetitions compared to the CRF, allowing better modeling of more complex disfluencies. We conjecture that the dense word representation in the BLSTM captures more of the reparandum/repair "rough copy" similarities than the simple POS pattern-match features.

6. Conclusion and Future Work

In summary, this paper introduces a Bidirectional LSTM neural network approach to disfluency detection, achieving state-of-the-art performance of 85.9 F-score on the standard disfluency detection task using explicit repair states, lexical feature embeddings, and integer linear programming post-processing. In addition, we improve the state-of-the-art in correction detection. Analysis shows that performance gain is for cases that are hardest to detect: restarts and repairs.

The best case BLSTM models leverage engineered pattern match features, indicating that the BLSTM architecture is not sufficiently powerful to learn these cues automatically with the amount of available annotated training data. While the pattern match features are known to be useful in cross-domain scenarios [1], an open question for future work is whether other neural network architectures might more effectively learn these cues. Another question is whether dynamic programming alternatives to ILP might improve performance. The experiments described here use hand transcripts, but the BLSTM framework is well-suited to combining text and prosodic features because of the continuous-space representation of text, which is another direction for future work.

7. References

[1] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Multidomain disfluency and repair detection," in *Proc. Interspeech*, 2014.

[2] M. Johnson and E. Charniak, "A tag-based noisy channel model of speech repairs," in *Proc. ACL*, 2004.

[3] M. Honnibal and M. Johnson, "Joint incremental disfluency detection and dependency parsing," *Transactions of the Association for Computational Linguistics*, vol. 2, no. 1, pp. 131–142, 2014.

[4] W. Wang, G. Tur, J. Zheng, and N. F. Ayan, "Automatic disfluency removal for improving spoken language translation," in *Proc. ICASSP*, 2010, pp. 5214–5217.

[5] E. Shriberg, "To errrris human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 01, pp. 153–169, 2001.

[6] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, Department of Psychology, University of California, Berkeley, CA, 1994.

[7] M. Ostendorf and S. Hahn, "A sequential repetition model for improved disfluency detection," in *Proc. Interspeech*, 2013.

[8] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1526–1540, 2006.

[9] K. Georgila, "Using integer linear programming for detecting speech disfluencies," in *Proc. NAACL HLT*, 2009.

[10] E. Cho, T.-L. Ha, and A. Waibel, "Crf-based disfluency detection using semantic features for german to english spoken language translation," in *Proc. IWSLT*, 2013.

[11] X. Qian and Y. Liu, "Disuency detection using multi-step stacked learning," in *Proc. NAACL HLT*, 2013.

[12] J. Ferguson, G. Durrett, and D. Klein, "Disfluency detection with a semi-markov model and prosodic features," in *Proc. NAACL HLT*, 2015.

[13] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proc. NAACL*, 2001, pp. 118–126.

[14] S. Zwarts, M. Johnson, and R. Dale, "Detecting speech repairs incrementally using a noisy channel approach," in *Proc. Coling*, 2010, pp. 1371–1378.

[15] M. S. Rasooli and J. R. Tetreault, "Joint parsing and disfluency detection in linear time." in *Proc. EMNLP*, 2013, pp. 124–129.

[16] J. Hough and D. Schlangen, "Recurrent neural networks for incremental disfluency detection," in *Proc. Interspeech*, 2015.

[17] E. Shriberg and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Proc. Eurospeech*, 1997, pp. 2383–2386.

[18] E. Shriberg, "Phonetic consequences of speech disfluency," in *Proc. International conference of Phonetics Sciences*, 1999, pp. 619–622.

[19] "Crf++," https://taku910.github.io/crfpp.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling." in *Proc. Interspeech*, 2012.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[23] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A field guide to dynamical recurrent neural networks*. IEEE Press, 2001.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[25] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[26] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ACL*, vol. I, 1992, pp. 517–520.

[27] "Theano," https://github.com/JonathanRaiman/theano_lstm.

[28] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.