Noisy BiLSTM-based Models for Disfluency Detection

Nguyen Bach and Fei Huang

Machine Intelligence Technology Lab, DAMO Academy, Alibaba Group

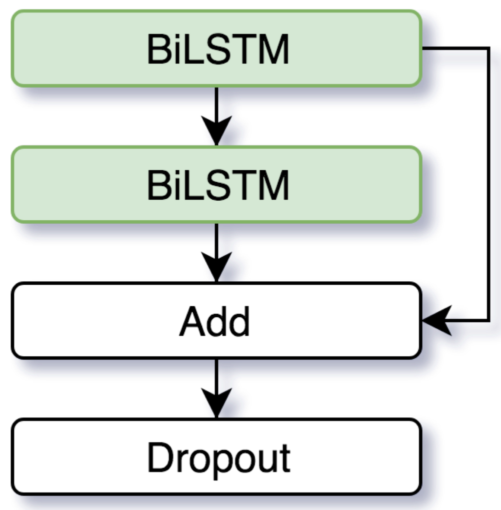
{nguyen.bach, f.huang}@alibaba-inc.org

Summary

- Disfluency detection based on residual BiLSTM blocks, self-attention, and noisy training
- Combining residual BiLSTM blocks, self-attention, and noisy training outperforms the BERT fine-tuned model by 1.2 F1 score on average across 4 non-Switchboard test sets.
- Our models are not only nearly 20 times smaller than BERT-based model but also surpasses BERT in 4 non-Switchboard test sets.

BiLSTM Residual

A ResNet-style BiLSTM residual block in which we skip one BiLSTM layer

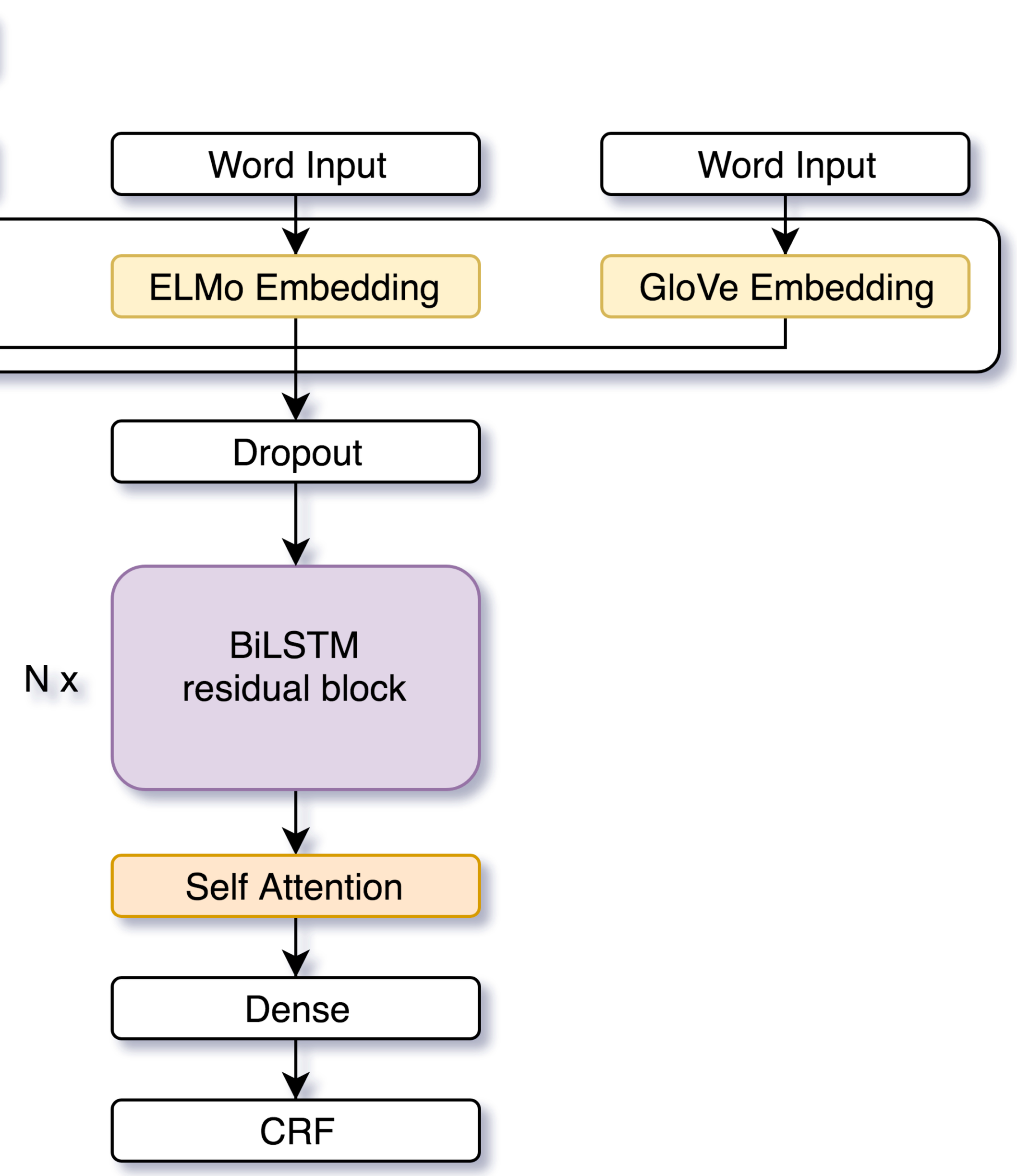


Noisy Training

- Insertion (iNoise): we first pick how many words to insert with a probability constraint. The next step is to randomly select insert position and words. All insertion words will be labelled as disfluency.
- Deletion (dNoise): we randomly delete a word in a given segment and after deletion all remaining words are labelled as ordinary, non-disfluent words.
- Repetition (rNoise): we randomly pick a position to start a repetition, and randomly pick repetition length in between 1 and 4 words. All repeated words are labelled as disfluency.
- We fix the percentage amount data which will be used during noisy training to 1 percent.

EGBC

EGBC is a residual BiLSTM-CRF network includes ELMo, GLoVe embeddings, character context LM, self-attention, and noisy training.



Training example on the Switchboard: *and/O i/O do/O n't/O have/O to/**@dis** to/O be/O writing/O checks/O*

Batch size and training epochs are set as 128 and 10 respectively. BiLSTM output space for word and character are 100 and 25 respectively. The residual BiLSTM contains 6 residual blocks.

Experimental Results

Table 1: *Results of disfluency detection on the English Switchboard data set. The first section shows pervious work. The second section describes our baseline with the richer word embedding method. The third section presents the contribution of residual, self-attention, and noisy training over our baseline. The forth section shows the BERT fine-tuned model performance.*

Method	**P**	**R**	**F1**
Weight sharing [17]	92.1	90.2	**91.1**
Transition-based [33]	91.1	84.1	87.5
BiLSTM [12]	91.6	80.3	85.9
Semi-CRF [11]	90.0	81.2	85.4
EGBC	95.9	86.3	90.9
GBC	93.1	80.9	86.6
BiLSTM CRF (BC) [20]	91.6	79.6	85.2
EGBC + residual + iNoise	95.7	88.3	**91.8**
EGBC + residual + self-attention	94.5	88.6	91.5
EGBC + residual	96.1	86.9	91.2
BERT fine-tune [1]	94.7	89.8	**92.2**

Table 2: *Results on 4 non-Switchboard test sets. The last column shows an average F1 score across 4 test sets. The first is our best model on Switchboard, and the second is the baseline model. The third section describes different noisy training scheme over the baseline including insertion, deletion, and repetition noise. The last section presents the model with residual BiLSTM block, self-attention, and insertion noise.*

Method	**CallHome**			**FCIC**			**SCOTUS**			**Interview**			**Average F1**
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
BERT fine-tune [1]	23.8	58.0	33.7	45.5	57.2	50.7	66.4	71.1	68.7	47.1	48	47.6	**50.18**
EGBC	24.8	55.9	34.3	46.3	54.8	50.2	67.1	71.4	69.2	53.5	43.5	48	50.43
EGBC + iNoise	23.9	60.8	34.3	48	57	52.3	69.4	70.6	**70**	47.9	44.9	46.4	**50.75**
EGBC + dNoise	22.6	59.5	32.8	43.1	58.6	49.7	63.4	70.7	66.8	45	43.9	44.4	48.43
EGBC + rNoise	21.4	66.5	32.4	40.2	63	49.1	62	72.6	66.9	41.7	47.5	44.4	48.2
EGBC + residual + self-attention + iNoise	24.7	60.1	**35.0**	50.7	53.8	**52.2**	71.7	67.7	69.6	57.5	42.3	**48.7**	**51.38**

Table 3: *The number of model parameters*

Method	Parameters
BERT fine-tune (base uncased)	110 million
EGBC	2.9 million
EGBC + residual + self-attention + iNoise	5.6 million

Examples

Table 4: Examples of disfluency detection. The red cross-out is our model disfluency detection. The blue underline is the reference annotation.

Switchboard	but they 've been in office since the ~~the~~ nineteen forties
CallHome	and then at night claudia gives ~~the~~ me a
FCIC	~~i~~ i not ~~referring~~ ~~to~~ ~~a~~ ~~specific~~ i referring to the fact that we look at our risks and we look at our positions
SCOTUS	~~what~~ ~~in~~ ~~in~~ what respect do you claim he is not properly deportable
Interview	life is a full of failures if people say ~~you~~ ~~know~~ nobody helped me when we do business you always want people to help you but if there 's no people to help you