Vol.2018-NL-238 No.4
2018/12/11

低頻度語学習手法を用いた Convolutional Encoder-Decoder モデルによる文法誤り訂正

町田 翔[1,a)]　藤田 和成[2]　延澤 志保[2]　荒井 秀一[2]

概要：文法誤り訂正は，英語学習者による誤った英文を自動で訂正するタスクである．近年，Recurrent Neural Network(RNN) モデルによる文法誤り訂正タスクの精度向上が報告されている．Neural Network モデルは，計算量削減のため学習時に扱えるボキャブラリに制限かけているという課題がある．英語学習者が誤る可能性の高い語は，出現頻度が低いためボキャブラリ数の制限により学習に含まれない．そのため，文法誤り訂正において低頻度語は，正しい語に訂正すべき対象である．これまで本研究では，文法誤り訂正のための低頻度語学習手法を提案し，RNN モデルを用いて精度向上を確認した．本稿ではさらなる精度向上を目指すため，広範的な文脈を補完することが可能な Convolutional Encoder-Decoder による文法誤り訂正のための低頻度語学習手法を提案する．我々は，データ数を増加させ，出現する単語数を増やさないために，限定的用法として使われている形容詞を削除するデータ拡張を行った．また，文字列としての出現頻度を底上げするために，単語を部分文字列として扱う低頻度語学習手法を行い，未知語の解消と文法誤り訂正タスクによる精度向上を確認した．

キーワード：文法誤り訂正，ニューラルネットワーク, 低頻度語学習手法，データ拡張．

Grammatical Error Correction on Convolutional Encoder-Decoder Model Focusing on a Learning Method of Low Frequency Words

Sho Machida[1,a)]　Kazumasa Fujita[2]　Shiho Hoshi Nobesawa[2]　Shuichi Arai[2]

Keywords: Grammatical Error Correction, Neural Network, Low Frequency word Learning Method, Data Augmentation.

1. はじめに

文法誤り訂正 (Grammatical Error Correction, GEC) は，英語学習者による誤った英作文を自動で訂正するタスクである．GEC タスクのコンペティションである CoNLL-2014 shared task では，動詞の時制や助動詞，前置詞などの 28 種類の誤りタグを定義している [1]．GEC において低頻度語は，英語学習者の誤る可能性が高い語が多く含まれるため，学習に含ませ訂正すべき対象である．近年，Neural Network を用いた分類器ベースによる GEC タスクの精度向上が報告されている [2], [3], [4], [5]．Neural Network モデルは，流暢性の高い文を出力できるが，計算量削減のため学習時に扱うボキャブラリサイズに制限を掛ける必要があるという課題がある．そのことにより低頻度語は，ボキャブラリサイズの制限により学習に含まれない．また，低頻度語を未知語（$unknown$ = <unk> タグ）として一律に扱ってしまっているため，流暢性や妥当性が掛けてしまうという問題点がある．そこで本研究では，低頻度語の出現頻度を底上げするため，データ拡張と単語を部分文字列に分割する低頻度語学習手法を行い，Neural Network モデルの課題の解決と精度向上を図る．我々は，Recurrent Neural Network(RNN) を用いた分類器ベースモデルを用いた GEC のための低頻度語学習手法を提案し，精度向上

[1] 東京都市大学大学院工学研究科情報工学専攻
Graduate School of Engineering, Tokyo City University
[2] 東京都市大学知識工学部情報科学科
Faculty of Knowledge Engineering, Tokyo City University
[a)] machida15@ipl.cs.tcu.ac.jp

を確認した [6]．さらなる精度向上と本手法の有効性を確認するため，本稿では広範的な文脈を補完することができる Convolutional Encoder-Decoder モデルによる GEC のための低頻度語学習手法を提案し，精度向上を図る．

2. 先行研究

Neural Machine Translation(NMT) は近年，RNN を用いた分類器ベースモデルによって精度向上が報告されている [7]．そこで，文法誤り訂正を誤った文から正しい文への翻訳として捉える NMT ベースの手法が提案されている [2]．このモデルを GEC に用いる場合，ボキャブラリサイズが大きいと decoder の学習が困難になるため，ボキャブラリサイズを制限する必要があり，低頻度語を校正することが困難であった．このように，ボキャブラリサイズの制限と training 中に出現していない単語を test 時に予測できないという問題点を解決するために，ボキャブラリをアルファベットで扱う手法では RNN を用いた分類器ベースモデルが提案された [3]．文字レベルでは，実質未知語を失くすことが可能であるが，文を構成する要素数が単語レベルと比べ多くなってしまう．そのため文字レベルでは，RNN のユニット数が増えてしまい，メモリ数との兼ね合いにより，長い文の学習と予測が困難であった．

GEC タスクのための Neural Network モデルは RNN に依存していた．そのため，より広範的な文脈及び離れた単語間の相互関係を補完するために Convolutional Encoder-Decoder モデルを用いる手法が提案された [5]．この手法は，RNN の手法より大幅に上回る精度を記録した．

本稿ではさらなる精度向上と本手法の汎用性を確認するため，Convolutional Encoder-Docoder モデルにおいても低頻度語学習手法が有効であることを示す．

3. 低頻度語学習手法

ここでは，training コーパスとして第二言語学習を支援するサービスである「語学学習 SNS lang-8」の lang-8[*1] を用いる．lang-8 は，英語学習者が作成した文 (source) と英語熟練者が校正した文 (target) で構成される会話文を収録したパラレルコーパスとは別に，すべての言語学習者と添削者の会話文を収録した大規模なパラレルコーパスがある．我々は言語識別ツール[*2] を用いて英語文以外を削除し，CONLL-2014 の training コーパス (NUCLE) と合わせて使用する．lang-8 は，約 27 種類の単語を含んでおり，出現回数が 10 回未満の事なり単語数が 90%を占めている．Neural Network モデルにおける GEC では，制限されたサイズのボキャブラリを用いて terget 文を生成するため，誤りを多く含む低頻度語を <unk> として統一に扱っている．単語レベルの RNN モデルでは，training コーパスに出現する約 27 万単語をボキャブラリサイズの制限により約 3 万単語にまで減らすため，約 24 万単語分の役割を <unk> が担うこととなる．<unk> と変換されてしまう約 24 万単語は，名詞や動詞，スペルミス語などの様々な語の役割を担うため，予測文に <unk> が含まれている場合，学習したモデルの信憑性が損なわれてしまう．ボキャブラリサイズの制限による予測文の例を図 1 に示す． 図 1 は，上が

図 1 ボキャブラリサイズの制限による予測文の例

入力文であり，下が出力文である．赤は低頻度語を意味する．ここでは，「frend」は「friend」のスペルミスのため低頻度語となり，「Vietnamese」はベトナム語の意味を持つが「Vinamese」とスペルミスをしている低頻度語であるため <unk> となって出力されている．図 1 に示すように予測文に <unk> が含まれることによって，誤り訂正すべき箇所を <unk> としてそのまま出力したり訂正不可能になったりしてしまう可能性が高い．そのため，モデルから生成された文の妥当性や，流暢性が失われてしまう．そこで我々は，低頻度語を Neural Network モデルの学習に含ませる手法を提案する．低頻度語学習手法は，前処理であるため 任意の Neural Network モデルに適応できる．さらに文法を崩すことなくデータ拡張することができることが本手法の利点である．

3.1 単語から部分文字列への変換

異なり単語数をボキャブラリサイズに収めるため，低頻度語を部分文字列として扱う Byte Pair Encoding(BPE)[8] を用いる．BPE は，単語を頻出する部分文字列に系列に分解することができるデータ圧縮手法である．例えば low，lower，love，loved の 4 語が低頻度語の場合，共通している「lo」とその他に分割し，分割した文字列の出現頻度が高くなるようにする効果がある．lang-8 の場合，出現頻度が 10 未満の異なり単語数の割合が約 90%を占めていたが，BPE を用いることにより異なり単語数の割合を抑えることが可能なる．また，BPE によって低頻度語を分割し文字列として大幅に出現頻度を向上させることが期待でき，単語ベースでは <unk> とされてしまう低頻度語を学習に含ませることが可能になる．しかし，BPE のみでは低頻度語を学習に含ませることが可能となるが，低頻度語が含まれる文が増加されるわけではない．そのため，本稿では文を増加させるデータ拡張を提案する．

3.2 データ拡張

Neural Network モデルは高い表現力の代わりに過学習

*1 言語学習 SNS lang-8：http://lang-8.com
*2 言語識別ツール：https://github.com/saffsd/langid.py

の恐れがあり，一般的に高い精度を得るためには大量の training データが必要になることが知られている．また，training コーパスに含まれる単語の約 90%は出現回数が 10 回未満の低頻度語である．そこで，我々は語の出現頻度を底上げすることによって，Neural Network モデルによる学習が困難である低頻度語の異なり単語数の削減を図る．また，低頻度語が出現する文を増加させ，モデルの表現力を高める必要がある．

GEC において target は，正しい文法で記述されている必要がある．そのため，関沢らの手法 [9] を参考に形容詞削除のデータ拡張を行う．本研究では形容詞の用法[*3]に着目する．形容詞の用法の例を表 1 に示す．表 1 のうち，叙述的用法と名詞的用法の形容詞を削除した場合に文として不完全な文となってしまう．GEC は英語学習者の文法誤りを訂正するタスクであり，添削者の文法は崩してはならない．そのため，我々は文法を崩さないデータ拡張を提案する．Python の NLTK パッケージ[*4]を用いて英文に品詞タグを付与し，限定的用法の形容詞を対象に削除した文を生成した．形容詞の後ろに名詞が連なっている場合，その形容詞を削除した文を生成していく．lang-8 に対してデータ拡張したことによって，生成した文の例を表 2 に示す．表 1 から，「限定」のみを対象として含まれる形容詞を削除した文の例である．1 文に形容詞が複数連なっている場合は，表 2 のようにデータ拡張する．「org」文には限定的用法の形容詞が 2 語含まれているため，片方を削除した文をそれぞれ作成し，双方とも削除した文を生成する．限定的用法に着目することにより文法を崩すことなくデータ拡張できていることがわかる．Lang-8 をデータ拡張したコーパスの低頻度語の変化を表 3 に示す．データ拡張の有効性を確認するため，表 3 に示す 3 つの拡張コーパスを作成した．表 3 のうち，「なし」は lang-8 の元コーパスを示し，「すべての形容詞」はすべての形容詞に対して形容詞削除のデータ拡張を行ったコーパスである．表 3 の「限定的用法」は限定的用法の形容詞に対して形容詞削除のデータ拡張を行ったコーパスを示し，「限定的用法（誤り文のみ）」は誤り文のみに限定的用法の形容詞削除のデータ拡張を行ったコーパスである．「誤り文のみ」に限定した理由は文法を学習させるだけでなく，誤りを訂正できるのではないかとの仮説の下に誤り文のみに対してデータ拡張を行ったためである．Neural Network モデルでは制限されたサイズのボキャブラリを用いて予測文を生成するため，10 回程度の出現頻度の単語はボキャブラリに含まれない．そのため，10 回未満の単語を低頻度語とする．「限定的用法」の場合，コーパスに出現する文が約 56 万文増加し，誤り文数の増加と約 13,000 単語の低頻度語の出現回数を向上させた．また，テストコーパスに出現する約 94%の単語の出現頻度の上昇も確認した．すべての形容詞に対して形容詞削除を行うことで限定的用法の約 2 倍のデータ拡張ができた．

表 1 形容詞の用法の例

用法	用例，解説
限定	Sumo is one of the *Japanese traditional* sports. 名詞の前に置かれ，名詞の意味を限定する働き
叙述	This watch is *waterproof* (= proof against water) 動詞のあとに置かれ補語となる場合
名詞	The *rich* (= Rich people) are apt to despise the *poor* (=poor people) . 形容詞が名詞として用いられる場合

表 2 データ拡張の例

org	Sumo is one of the *Japanese traditional* sports .
DA1	Sumo is one of the *traditional* sports .
DA2	Sumo is one of the *Japanese* sports .
DA3	Sumo is one of the sports .

表 3 データ拡張によるコーパスの変化

データ拡張	DA により増加した文数	誤り文の比率	低頻度語の比率
なし	-	57%	90%
すべての形容詞	1,128,718 文増	64%	80%
限定的用法	565,439 文増	59%	85%
限定的用法（誤り文のみ）	355,982 文増	64%	83%

4. 実験と評価

低頻度語学習手法の有効性を確認するため，大きく分けて以下の 3 つの実験を行った．

実験 1 データ拡張の有効性を確認するため，データ拡張の有無で実験し評価値を比較する．

実験 2 低頻度語学習手法の汎用性を確認するため，RNN モデルと Conolutional Encoder-Decoder モデルの 2 つのモデルで実験し，双方での精度向上を確認する．

実験 3 さらなる低頻度語手法の有効性を確認するため，既存研究との評価値を比較する．

これらの実験から，広範的な文脈を補完することが可能な Convolutional Encoder-Decoder モデルと低頻度語学習手法の組み合わせが有効であることを示す．GEC のコンペティションである CoNLL-2014 の指標には，再現率よりも適合率を重視するため $F_{0.5}$ 値が用いられている [1]．低頻度語学習手法を用いた場合のコーパスの異なり単語数と出現頻度の関係を表 4 に示す．表 4 の lang-8 は元コーパスを示し，低頻度語学習手法は lang-8 に対して低頻度語処理を行った場合のコーパスであり，要素は異なり単語数を示す．低頻度語は出現回数が 10 回未満の単語を示す．表 4 から，低頻度語学習手法により全体的に異なり単語数を減少させることができ，低頻度語が約 99%削減された．また，100 回以上出現する語の異なり単語数が lang-8 より増加していることがわかる．限定的用法の形容詞削除のデータ拡張を

*3 形容詞の用法：『ジーニアス英和大辞』小西 友七，南出康世編集主幹，大修館 2001
*4 NLTK パッケージ：https://www.nltk.org/

表 4　低頻度語学習手法による異なり単語数と出現頻度の変化

出現頻度	lang-8	lang-8（低頻度語学習手法）
低頻度語	236,819 語	4,257 語
10 回以上 50 回未満	22,110 語	6,364 語
50 回以上 100 回未満	4,319 語	9,119 語
100 回以上 500 回未満	5,296 語	10,056 語
500 回以上 1,000 回未満	1,129 語	1,807 語
1,000 回以上	1,852 語	2,714 語

することによって，テストコーパスに出現する約94%の単語の出現頻度の向上を確認した．データ拡張によってどのような評価値に影響がでるのかを検証していく．

4.1　データ拡張の有無による実験比較

限定的用法に着目した形容詞削除のデータ拡張が有効性を確認するため，データ拡張の有無による実験結果を表5示す．表5に示すようにモデルを固定し，4つのコーパス

表 5　データ拡張によるコーパスの変化

モデル	データ拡張	$F_{0.5}$ 値
CNN-BiLSTM	なし	49.77
	すべての形容詞削除	49.24
	限定的用法の形容詞削除	51.06
	限定的用法の形容詞削除 (誤り文のみ)	50.04

に対して実験を行った．すべての形容詞に対して形容詞削除のデータ拡張を施してしまうと，評価値が下がってしまい，限定的用法に対して形容詞削除のデータ拡張を行うことにより評価値が上昇することがわかった．また，誤り文のみに対してデータ拡張を行うと評価値が限定的用法の形容詞削除と比べて評価値が下がることがわかった．データ拡張なしより，限定的用法に制限をかけたデータ拡張の方が1.29point向上することがわかった．

データ拡張なしの場合と限定的用法に対してデータ拡張した場合の差が数pointしか離れていないため，予測文の違いを考察する．予測文の実際の例を図2に示す．図2の

図 2　データ拡張の有無による予測文の違い

紫が誤りを示し，赤は訂正語を意味する．四角は文字列の削除を意味する．緑は低頻度語を示す．データ拡張なしでは，「reimbursement（：代償，払い戻し）」は8回しか出現しない低頻度語であったが，データ拡張することによって116回出現させることができた．そのことにより，「for ＊ reimbersement」という言い回しを学習できた．また，「leaks」の単語3-gramのパターンを見てみると，「leaks from ＊」という言い回しがlang-8コーパスに多く出現するためにデータ拡張なしでは誤り訂正できなかった．しかし，データ拡張を行うことによって動詞の誤りを捉えることができた．

4.2　低頻度語学習手法を用いたモデルの比較実験

低頻度語学習手法の汎用性を確認するため，RNNモデルとConolutional Encoder-Decoderモデルの2つのモデルで実験し，双方でも精度向上するのかを確認する．図3のRNNモデルは，sourceを文字列ごとにencodeし，attention mechanismを介してdecoderで正解文を文字列ごと予測するアーキテクチャである．RNNモデルはZiangらのモデル[3]を参考にした．target word vectorは予測する文字列の教師であり，attentionで得たベクトルと共にdecoderに入力し学習する．@タグはBPEによる分割を表す．encoderをBidirectional Long Short Term Memory(Bi-LSTM)[10]の3層とし，decoderを2層に設定した．図の4のConvolutional Encoder-Decoderモデルは，CNNのEncoderで広範的な文脈の特徴を捉え，Bi-LSTMのDecoderで予測文を生成するアーキテクチャである．CNN-BiLSMTモデルは，Chollampattら[5]を参考にした．ボキャブラリサイズは1万，2万，3万の3種類で実験し，$F_{0.5}$ が最も高く

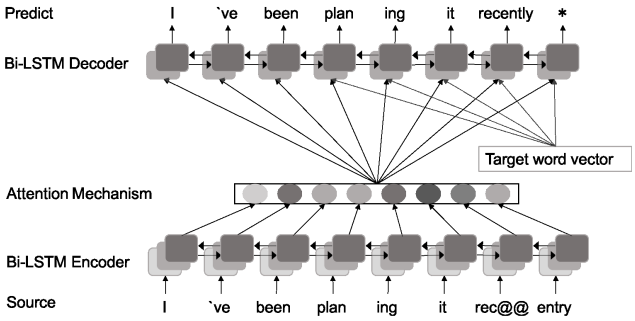

図 3　RNN モデルのアーキテクチャ

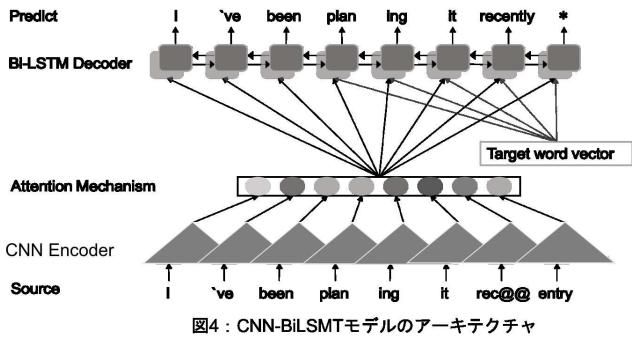

図 4　Convolutional Encoder-Decoder モデルのアーキテクチャ

なった 2 万をボキャブラリサイズとして選択した．2 つのモデルの Embedding サイズや扱う文の長さなどは同じパラメータを設定した．CoNLL-2014 shard Task[1] に従って test した．2 つのモデルに対して低頻度語学習手法を用いた実験を行い，双方でも精度向上するかを評価比較する．結果を表 6 に示す． 表 6 に示す通り，RNN モデルと

表 6 低頻度語学習手法を用いたモデルの比較

モデル	低頻度語学習手法	$F_{0.5}$ 値
RNN	なし	43.85
	あり	45.36
Convolutional Encoder-Decoder	なし	49.77
	あり	51.06

Convolutional Encoder-Decoder モデルの双方とも精度向上が見られた．低頻度語学習手法はモデルに依存することなく，有効性があることがわかった．

4.3 既存研究との評価比較

本稿では，実験 2 で評価値が高かった Convolutional Encoder-Decoder モデルを使用し既存研究との評価値を比較し，表 7 に示す． 表 7 のうち単語ベースモデルでは，独

表 7 低頻度語学習手法を用いたモデルの比較

モデル	コーパス	備考	$F_{0.5}$
単語ベース [2]	lang-8	非公開コーパスを使用	40.56
文字ベース [3]	lang-8 + NUCLE		40.58
Hybrid モデル [4]	lang-8 + NUCLE	単語埋め込みを使用	45.15
Convolutional Encoder-Decoder[5]	lang-8 + NUCLE	単語埋め込みを不使用	50.70
Convolutional Encoder-Decoder	lang-8 + NUCLE	低頻度語学習手法	**51.06**

自の非公開コーパスを使用している．Hybrid モデルは単語埋め込みを使用しているため，文字ベース，単語ベースと比較すると大幅に評価値が向上していることがわかる．表 7 から提案手法は，ボキャブラリの制限がない文字ベースの手法 [3] と比べても，精度が良いことがわかった．従来手法と比較した結果．データ拡張と単語を部分文字列として扱い，低頻度語を学習に含ませることにより，GCE タスクにおいて精度向上したことから，提案手法の有効性を確認した．

低頻度語学習手法を用いることによる流暢性や妥当性の向上を考察する．実際の予測文を図 5 に示す．図 5 から，低頻度語学習手法を用いることによって <unk> タグの解消をすることができていることがわかる．低頻度語をボキャブラリに含ませ，低頻度語の出現する文を増加させることにより <unk> タグとして扱っていた単語を誤り訂正することができた．

5. おわりに

本稿で提案した低頻度語学習手法により，GEC タスクにおけるさらなる精度向上を確認した．単語の出現回数を底上げし，文字列として扱うことにより，誤った語が多く含

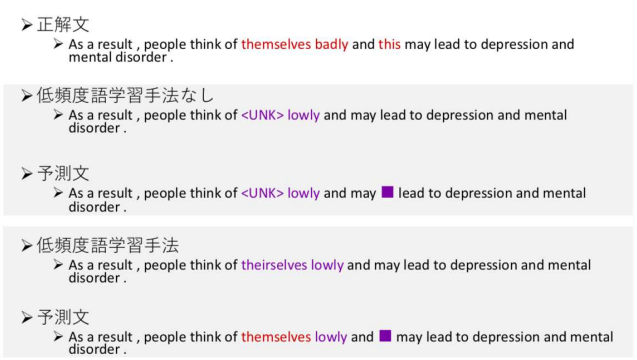

図 5 <unk> タグの誤り訂正の例

まれる低頻度語を学習に含ませることを可能とした．英語学習者の誤りに形容詞誤りが少ないため，データ拡張に形容詞削除を用いることが効果的であったと考えられる．すべての形容詞に対して形容詞削除を行うよりも，限定的用法に制限したデータ拡張の方が文法を学習することができることがわかった．形容詞の限定的用法に制限したデータ拡張により，時制や文法を学習することができ，評価値は 51.06point を記録した．また，低頻度語学習手法はモデルに依存せず，汎用性があることがわかった．提案手法を既存手法と比較し，0.3point の向上を確認した．さらなる精度向上を目指すため，Convolutional Encoder-Decoder モデルの改良をする．

参考文献

[1] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. and Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction, *Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14 (2014).

[2] Chollampatt, S., Taghipour, K. and Ng, H. T.: Neural Network Translation Models for Grammatical Error Correction, *Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 2768–2774 (2016).

[3] Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D. and Ng, A. Y.: Neural Language Correction with Character-Based Attention, *CoRR*, Vol. abs/1603.09727, (2016).

[4] Ji, J., Wang, Q., Toutanova, K., Gong, Y., Truong, S. and Gao, J.: A Nested Attention Neural Hybrid Model for Grammatical Error Correction, *CoRR*, Vol. abs/1707.02026, (2017).

[5] Chollampatt, S. and Ng, H. T.: A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction (2018).

[6] 町田翔, 延澤志保, 荒井秀一：ニューラル文法誤り訂正モデルにおける低頻度語処理法の提案, FIT2018（第 17 回情報科学技術フォーラム）, 第 2 分冊, No. E-020, pp. 187–188 (2018).

[7] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *CoRR*, Vol. abs/1409.3215, (2014).

[8] Gage, P.: A New Algorithm for Data Compression, *C Users J.*, Vol. 12, No. 2, pp. 23–38 (1994).

[9] 関沢祐樹, 梶原智之, 小町守：目的言語の低頻度語の高頻度語への言い換えによるニューラル機械翻訳の改善, 言語処理学会 第 23 回年次大会 発表論文集, pp. 982–985 (2017).

[10] Schuster, M. and Paliwal, K.: Bidirectional Recurrent Neural Networks, *Trans. Sig. Proc.*, Vol. 45, No. 11, pp. 2673–2681 (1997).