The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Bidirectional LSTM を用いた誤字脱字検出システム

Misspelling Detection by using Multiple Bidirectional LSTM Networks

高橋 諒[*1] 蓑田 和麻[*1] 舛田 明寛[*2] 石川 信行[*1]
Ryo Takahashi Kazuma Minoda Akihiro Masuda Nobuyuki Ishikawa

[*1] 株式会社リクルートテクノロジーズ Recruit Technologies Co.,Ltd.
[*2] 株式会社 PE-BANK PE-BANK, Inc.

Companies in the RECRUIT Group provide matching business between clients and customers, and create lots of manuscripts every day in order to tell the attractiveness of our clients. In this paper, we propose a method for detecting misspelling in manuscripts with machine learning. That system mainly consists of two parts. One is the multiple Bidirectional LSTM networks to estimate the probabilities of correctness in each characters. The other is the random forests algorithm to decide what sentence is correct or not by using outputs of these networks. The efficacy of our approach is demonstrated on two datasets: artificial sentences and real manuscripts created in our services.

1. 背景・課題

情報を提供するクライアントと情報を求めるカスタマーをマッチングするのがリクルートのビジネスモデルである. このビジネスモデルにおいて, 原稿はクライアントとカスタマーを結びつける重要な情報伝達手段である. その原稿において, 万が一誤った内容が記載されてしまうと, 企業としての信頼に関わる問題に発展するため, 校閲業務に対しては多くのリソースが割かれている. しかし, それでも稀に不備のある原稿が発見されたり, 文章として違和感のある原稿が掲載される事態が発生したりすることが現状であった. その原因の 1 つが, 1 枚の原稿に対してチェックすべき規定の多さである. 通常どのサービスにおいても, それぞれに独自の原稿規定が存在しており, その数は各々100~200 項目に渡る. 年間数十万件の規模で新規原稿が作成される中で, それら 1 枚 1 枚に対し全項目のチェックを人手で行うのは困難であり, 熟練した校閲者でも見落としが発生していた. 特に, 多くの規定の中でも, 誤字脱字についてはチェックが十分に行われていないケースが散見されていた.

上記の課題に対し, システムによる校閲補助および自動校閲が出来ないかの検討を行った. 具体的には, NG ワードチェックのようなルールベースでの校閲に加え, 誤字脱字や原稿内での表記ゆれの検出を機械学習により行うハイブリッドな校閲システムを作成し, 実導入を行った. 本論文においては, その中でも深層学習を用いた誤字脱字検出ロジックについてのアルゴリズムとその実験/導入結果について記す.

誤字脱字の典型例として, ”私は猫が空きです”のような漢字の変換ミスや”私に猫がに好きです”というような助詞の間違いなどがある. このような誤字脱字は, 単語の組み合わせで検出しようとすると, その数が膨大になり全てを定義することは困難なため, ルールベースによる検出は難しい. そこで, 本研究では文字の系列情報を利用できる深層学習を利用したアプローチを試みた.

2. 関連研究

深層学習の分野において, 様々なネットワークが提案されているが, 文章や音声波形などの系列データに対して有効なネットワークとして Recurrent Neural Network（RNN）[1] が存在する.

連絡先：株式会社リクルートテクノロジーズ
IT エンジニアリング本部データテクノロジーラボ部 高橋 諒
（ryo_takahashi@r.recruit.co.jp）

自然言語処理という観点で, RNN を用いた事例として代表的なものとして文章の自動生成がある[2][3]. RNNを生成モデルとして捉え, 文章として成立している文字列 $(x_1, \dots, x_T)$を入力とし, それぞれの次の文字を示す$(x_2, \dots, x_{T+1})$を正解として学習させることで, 時刻 t+1 に出現する文字の確率$P(x_{t+1}|x_1, \dots, x_t) = softmax(o_t)$ を取得する. ここでo_tとは、時刻 t におけるネットワークの最終出力である。この確率値 $P(x_{t+1}|x_1, \dots, x_t)$が最大となる文字を次の文字として順々に生成することで文章の生成を行う.

本研究においては, この言語モデルとしての RNN をベースに異常検知として利用している. RNN を異常検知の文脈で利用した研究として, 例えば[4]や[5]が存在する. いずれも RNN の最終層に二値分類を行うための Dense Layer を繋ぐことで, その系列が正常か異常かを判定している. この形式の異常検知アルゴリズムを, 日本語の言語処理に適用した先行研究として[6]が挙げられる. [6]ではテレビで利用されるテロップにおける誤字脱字検出を目的とし, 誤字脱字を予め 8 つのパターンに分類し, それぞれのパターンに対して正常/異常の二値分類を行う RNN モデルの構築を行っている. [6]ではパターン毎のモデルの結果を単一文章に対して重ね掛けで検出した場合, モデル数が増えると精度が下がる点が指摘されている.

提案手法では, RNN の中でも, 長期依存性をもつ LSTM[7]を双方向に発展させた Bidirectional-LSTM（BLSTM）[8]を採用した. 更に, 言語モデルと正常/異常の二値分類の BLSTMを並列で利用し, それぞれの出力値を入力としたランダムフォレストの結果から誤字脱字を含むか否かの判定を行った. 提案手法の特徴は下記 4 点である.

・BLSTM を利用することでターゲットとなる文字の前後双方の情報を利用できる
・言語モデルを組み合わせているため, 予め考えられない誤字脱字のパターンに対しても対応ができる
・言語モデルの出力結果を参照することで, 誤字脱字判定された文字の代替提案が可能である
・複数の BLSTM のモデルの出力の組み合わせにランダムフォレストを利用することで, 検出時の閾値設定が容易になった

3 章では用いたデータセットについて, 4 章では提案手法の詳細について述べ, 5, 6 章では実際のデータを用いた実験と

試験運用結果について紹介し，7 章で今後の展望について説明を行う．

3. 学習データセット

本手法では，BLSTM/ランダムフォレストモデルの学習用に誤字脱字を含まない文(OK 文)と含む文(NG 文)のデータセットが必要である．リクルートには校閲済み原稿が大量に存在するため，これを OK 文として利用する．次に，過去の校閲内容の分析より頻出の誤字脱字のパターン(以下，NG パターン)を表 1 のように定義し，OK 文を基にして NG パターンに該当する誤字脱字を含む NG 文を作成することにした．

対象原稿としては後述する試験運用を見据え，リクルートが運営するサービスの一つであるゼクシィの原稿を利用した．

用意した学習データセットを表 2 に示す．校閲済み過去原稿は 2015 年 1 月～2018 年 1 月に掲載された原稿であり OK 文のみで構成される．作成 NG 文Ⅰ/Ⅱは NG 文と，その基となった OK 文のペアで構成され，誤字脱字の箇所の情報も含む．なお，作成 NG 文Ⅰは人手で作成したが，そこで不足した NG パターンを補うため作成 NG 文Ⅱを機械的に作成した．

表 1: NG パターン定義

名称	内容	例
漢字	漢字変換ミス	正）100名まで収容可能な会場。 誤）100名まで収容可能な海上。
助詞連続	助詞の不自然な連続	正）ドレスのご試着は、 誤）ドレスのをご試着は、
脱字(送り仮名)	送り仮名の脱字	正）ご要望にお応えします。 誤）ご要望にお応します。
脱字(助詞)	助詞の脱字	正）写真撮影を行います。 誤）写真撮影行います。
英字混入	タイプミスなどによる英字混入	正）宜しくお願いします。 誤）宜しくお願いしまs。

表 2: 学習データセット

データセット名	含まれるNGパターン	NG文作成方法	文数[件] ※1	データ量[MB]
校閲済み過去原稿	-	-	428,716	46
作成NG文Ⅰ	漢字、助詞連続、脱字(送り仮名)	クラウドソーシングを利用して人手で作成	36,565	10
作成NG文Ⅱ	脱字(送り仮名)、脱字(助詞)	プログラムによる自動生成．平仮名をランダムに選んで除去	1,247,690	291

※1 原稿を句点などの終端文字で区切った 1 文を 1 件とする．ただし，作成NG文Ⅰ，Ⅱについてはok/ng文のペア数を表記．

4. アルゴリズム説明

本章では深層学習を用いた誤字脱字検出ロジックについて説明する．方針として，まず文字ごとの妥当性を判断する BLSTM モデルを構築し，その出力から文単位での正誤を判断するランダムフォレストを構築する．これらを組み合わせ，最終的には「誤字脱字箇所」，「正しい候補の文字」，「誤字脱字を含む文」の3つを出力する．

4.1 BLSTM による文字毎の OK/NG 確率モデル

本手法では，前方向からの文字の流れだけでなく，後方からの情報も捉えることができる BLSTM を採用した．文字毎の正常/異常を求めるニューラルネットワークのアーキテクチャーを図 1 に示す．BLSTM による出力は文字毎に順方向/逆方向の 2 つが存在するため，それらを結合し，各文字が正しいまたは誤字脱字である確率(OK/NG 確率)の 2 次元を出力するよう設計した．BLSTM 部分は順方向/逆方向で 2 層ずつ，計 4 層の中間層を持つ設計とした．損失関数にはクロスエントロピーを用い，文字毎に誤差を足しあわせた値を 1 文の誤差と定めた．推論の際は，上記の枠組みで学習されたモデルを用いて各文字の OK/NG 確率を出力し，この出力結果を利用して「誤字脱字箇所」と「誤字脱字を含む文」を判定する．

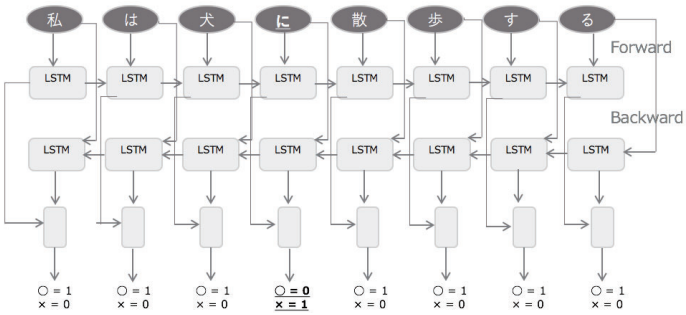


図 1: BLSTM による文字毎の OK/NG 確率モデル

4.2 BLSTM による言語モデル

BLSTM による言語モデルのアーキテクチャーを図 2 に示す．入力部分は基本的に図 1 と同様である．異なる点は以下 3 つである．

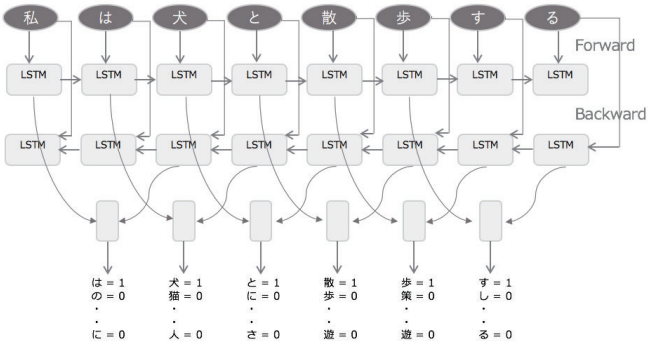


図 2: BLSTM による言語モデル

1. 学習に用いるデータの種類

学習に用いるデータは OK 文のみとした．言語モデルは正しい文から正しい文字の流れを予測するためである．

2. 最終層への入力を作成するロジック部分

4.1 の OK/NG 確率モデルとの違いは，最終層への入力を作成するロジック部分である．図 2 に示すように言語モデルの場合，t 番目の文字を予測するために，順方向 LSTM の t-1 番目の文字における出力値と，逆方向 LSTM の t+1 番目の文字における出力値を結合し，最終層への入力とする．言語モデルの場合，前後の文字から該当文字を予測するため，このような設計とした．

3. 最終層の出力

最終層の出力は基本的な言語モデル同様，文字サイズ分の次元を出力する設計である．

推論の際は，上記の枠組みで学習されたモデルを用いて，文を構成している文字に対する確率を出力し，この出力結果を利用して「誤字脱字箇所」と「誤字脱字を含む文」，「正しい候補の文字」を判定する．この言語モデルを用いる利点は，予想できていない誤字脱字を検出できる可能性がある点である．4.1 の

モデルのみでは機械的に作られた NG データを基にしているため，予想していない誤字脱字の検出力が弱くなる．それに比べ，言語モデルでは日本語として正しい文字の流れを学習するため，予想していない誤字脱字を検出できる可能性が高くなる．

4.3 複数モデルのアンサンブル方法

4.1, 4.2 で作成したモデルの出力値を使用して，入力文に誤字脱字を含むかどうかを判定する．使った変数を表 3 に示す．

表 3: ランダムフォレストの入力変数表

変数の説明	次元数
a) BLSTMモデルの出力する確率が最低となる箇所	1
b) a)における文字の確率	1
c) a)における文字の種別 ※1	6

※1: {平仮名,カタカナ,漢字,英字,数字,その他}のいずれかを示すone hot vector.

BLSTM モデル毎に上記を求め，全モデル分統合したものを入力とし，「誤字脱字を含む文か否か」の 2 値を分類するモデルを作成した．学習器にはランダムフォレストを使用した．推論時はそのランダムフォレストが出力する確率値と閾値の比較により判定する．このように複数の BLSTM モデルの出力値を用いた学習器を使用する事で，各々の BLSTM モデルの出力値に対する閾値をチューニングする必要がなくなる．さらに誤字脱字と判定する際の基準の選定の精度向上にも繋がり，精度面/保守面共に良いパフォーマンスとなる．

4.4 誤字脱字箇所推定と候補文字の決定

入力文に対しランダムフォレストが「誤字脱字を含む文」と判断した場合，「誤字脱字箇所の推定」と「候補文字の決定」を行う．誤字脱字箇所の推定は，各 BLSTM モデルの文字毎の確率が一定閾値以下となった箇所とする．誤字脱字箇所と推定された部分に対しては候補文字を決定する．候補文字は誤字脱字箇所において BLSTM 言語モデルの出力する確率が高い上位 3 文字とする．ただし，余分な文字が入っている，または脱字のような NG 文は，誤字脱字箇所を候補文字で置き換えるだけでは文の修正ができない点に注意が必要であり，今後の課題とする．

5. 実験

5.1 文単位の性能評価

評価に用いるモデルを表 4 に示す．LSTM 言語モデルは前方から後方へ向かう LSTM のみで構成した言語モデルである．BLSTM（言語，OK/NG 確率Ⅰ，Ⅱ）モデルは 4.1, 4.2 で述べたモデル，アンサンブルモデルは BLSTM と 4.3 で述べたランダムフォレストで構成されるモデルを指す．BLSTM OK/NG 確率モデルは，学習データセットである作成 NG 文ⅠとⅡでサイズや内容が違うため，それぞれでモデルを分けた．

評価用データとして OK 文と NG 文を同数用意し，各モデルでの NG 文に対する検出率(True Positive Rate)と OK 文に対する誤検出率(False Positive Rate)で評価する．NG 文は表 1 に示す NG パターンごとに 200 文ずつ作成した．ここでは，句点等の終端記号で区切った単位を1文とし，NG 文1文あたり1つの誤字脱字を含むようにした．各モデルは1文ごとに誤字脱字を含む/含まないを判定する．アンサンブル以外のモデルは文字ごとに正しさ表す確率を出力するため，文に含まれるの全文字の確率最低値と閾値との比較で判定する．アンサンブルモデルの出力は文単位での確率であるため，出力値と閾値の比較で判定する．

ROC 曲線と AUC を図 3，表 5 に示す．表 5 より，言語モデル同士で LSTM と BLSTM を比較すると脱字（助詞）を除く全 NG パターンで BLSTM の方が AUC 値で上回っている．BLSTM モデル同士（言語，OK/NG 確率Ⅰ，Ⅱ）の比較では，それぞれ得意な NG パターンが異なる．OK/NG 確率Ⅰ，Ⅱモデルは学習データに含まれる NG パターンに対して強く，その他の NG パターンに対して弱い．一方，言語モデルは漢字，英字混入に強い．特に英字混入は OK/NG 確率Ⅰ，Ⅱモデルの学習データに無い NG パターンであり，言語モデルの導入により未知の NG パターンに対応できる可能性がある．アンサンブルモデルは全体，漢字，助詞連続，脱字（送り仮名）について AUC が全モデル中最高値であり，他の NG パターンでも一定値を保っており，3つのモデルを統合することで相補的な効果が得られている．

5.2 誤字脱字箇所推定と候補文字の評価

誤字脱字箇所推定と候補文字について評価結果を表 6 に示す．また，検出文の事例を表 7 に示す．評価用のモデルはアンサンブルモデルを採用し，NG 文検出の閾値は誤検出率=0.200 となる値を採用した．このとき検出率=0.795 である．

表 6 より，誤字脱字箇所推定の正解率は 90.6%と高い．候補文字の正解率は 62.5%である．BLSTM 言語モデルは注目する箇所の前後の文字を正として利用しているため，表 7 No.2 のような 1 文字間違いのケースでは正解率が高いが，No.1 のような連続する 2 文字が間違うケースでは正解できないケースが多く見られた．

6. 試験運用

ゼクシィを対象に試験運用を実施した．ゼクシィ原稿の校閲者は校閲システムを利用し PC 画面上で確認や修正を行う．アンサンブルモデルを用いた誤字脱字検出ロジックを校閲システムに組み込み，試験運用した．試験期間中に投稿された原稿に対して校閲前後の文とアルゴリズムの検出結果を収集し，投稿された原稿を人手で OK 文と NG 文に振り分け，1 文単位での検出率/誤検出率で評価した．ただし，収集した NG 文には表 1 で定義していない NG パターンも含まれる．試験運用は期間を2期に分け，それぞれで評価した．評価結果を表 8 に示す．また検出できた事例を表 7 に示す．

表 8 より，第1期の検出率は 60%，誤検出率は 11%である．試験運用と評価用データでは NG パターンの分布に差があると考えられるが，第1期では図 3(e)の ROC 曲線と比較しても妥当な結果となった．その一方で適合率(Precision)は 14%と低い．理由として NG 文に対して OK 文は 29 倍と多いことが挙げられる．

適合率は校閲者にとってシステムに対する心理的な信頼度に直結するため向上のための対策が必要である．学習用データセットは試験運用の半年前までに取得したものであるが，原稿の文章は時間経過に伴うトレンドの変化により徐々に変わっていくと考えられる．そこで，直近のデータを使用すれば適合率の向上が期待出来ると考え，第1期で収集した原稿を学習データとして新たに追加し，作成済みの BLSTM モデルをファインチューニングした．

表 8 に示す第2期はファインチューニング後のモデルを適用した結果である．第1期と比べ誤検出率が 6%まで低下し，適合率が 22%まで上昇することが確認できた．

7. まとめ

本研究では，複数の BLSTM モデルのアンサンブルによる誤字脱字検出システムの開発および実験を行った．先行研究と比較して，提案手法では，OK/NG 確率モデルに加え言語モデルを組み込むことで，想定していない誤字脱字パターンの検出が可能になり，検出後の候補文字の提案まで可能となった．

実験では，リクルートが保有するサービスの実データを利用して学習を行い，試験運用を行った．その結果，誤字脱字のない OK データが圧倒的に多数を占める状態のなかで誤検出 6%，適合率 22%という結果を得た．

現時点での課題として，脱字のような単純に文字置き換えでは対応できないパターンでの候補文字の提案手法の確立や検出精度向上のためのネットワーク構造の見直しがある．また，運用面では，実際にシステムを利用してもらうことで蓄積されるフィードバックデータを順次追加で学習をしていく仕組みの構築を行う．

参考文献

[1] LWilliams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E. Learning representations by back-propagating errors, 1986

[2] Ilya Sutskever,James Martens,Geoffrey Hinton. Generating Text with Recurrent Neural Networks. 2011.

[3] Tsung-Hsien Wen,.Milica Gasi˘ c, Nikola Mrk ´ si˘ c,´Pei-Hao Su, David Vandyke,Steve Young Semantically Conditioned LSTM-based Natural Language Generation forSpoken Dialogue Systems. 2015.

[4] BBenjamin J. RadfordLeonardo M. Apolonio, Antonio J. Trias, Jim A. Simpson. Network Traffic Anomaly Detection Using Recurrent Neural Networks. 2018.

[5] Pankaj Malhotra1, Lovekesh Vig2, Gautam Shroff1, Puneet Agarwal. Long Short Term Memory Networks for Anomaly Detection in Time Series. 2015.

[6] 中野 信. AI 技術を使った誤テロップ自動検出に関する技術検証. The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018.

[7] Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, Vol. 9, No. 8, pp. 1735–1780, 1997.

[8] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, Vol. 45, No. 11, pp. 2673–2681, 1997.

表 4: 評価モデル

モデル名	内容	使用学習データ
LSTM 言語モデル	単方向LSTMによる言語モデル	校閲済み過去原稿
BLSTM 言語モデル	Bidrectional LSTMによる言語モデル	校閲済み過去原稿
BLSTM OK確率モデルⅠ	Bidrectional LSTMによるOK/NG確率モデル	作成NG文Ⅰ
BLSTM OK確率モデルⅡ	Bidrectional LSTMによるOK/NG確率モデル	作成NG文Ⅱ
アンサンブルモデル	上記３種類のBLSTMモデルとランダムフォレストで構成	作成NG文Ⅰ＋Ⅱ

表 5: 文単位性能評価結果(AUC)

モデル名	全体	漢字	助詞連続	脱字（送り仮名）	脱字（助詞）	英字混入
LSTM 言語モデル	0.79	0.89	0.59	0.8	0.7	0.95
BLSTM 言語モデル	0.83	0.96	0.62	0.88	0.62	0.99
BLSTM OK/NG確率モデルⅠ	0.76	0.88	0.89	0.95	0.51	0.58
BLSTM OK/NG確率モデルⅡ	0.77	0.59	0.62	0.94	0.87	0.77
アンサンブルモデル	0.88	0.97	0.89	0.96	0.77	0.88

黄色は特定NGパターンについて他モデルと比べ高い箇所

表 6: 指摘箇所推定と候補文字評価結果

NG文数	NG文 検出数	誤字脱字箇所 推定正解数 ※1	誤字脱字箇所推定正解率 ※1	候補文字評価 対象数 ※2	候補文字 正解数 ※2	候補文字 正解率 ※2
1000	795	720	0.906	376	235	0.625

※1 ： 推定した誤字脱字箇所中に真の誤字脱字箇所を含むとき正解とみなす.
※2 ： 候補文字３文字中に正しい文字を含むとき正解とみなす. ただし4.4節で述べた制約があるため、NG箇所の修正前後で文字数が等しい文のみ対象とする.

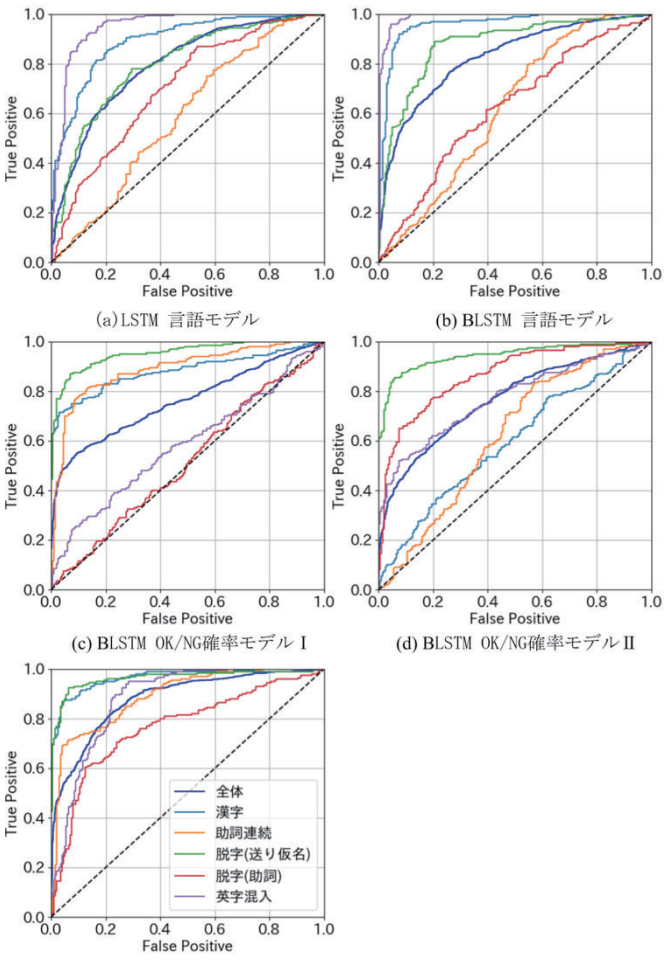
表 7: 検出成功事例

No.	評価データ	NG文と推定誤字脱字箇所	OK文	候補文字
1	作成NG文	フェアに【傘】【下】して結婚式のイメージを膨らませてみて！	フェアに参加して結婚式のイメージを膨らませてみて！	[参, 面, 開], [を, と, そ]
2		この機会にシェフ渾身のお料理【w】ご堪能ください。	この機会にシェフ渾身のお料理をご堪能ください。	[を, で, も]
3	試験運用時原稿	より格調高いしつら【れ】に変わり、いっそう厳かな雰囲気になる神殿。	より格調高いしつらえに変わり、いっそう厳かな雰囲気になる神殿。	[え, い, う]
4		海外にいるかのような4【つ】【パ】ーティ空間	海外にいるかのような4つのパーティ空間	[つ階月], [パ, ホ, の]

【】は推定誤字脱字箇所. 候補文字は【】１つあたり３文

表 8: 試験運用結果

試験運用フェーズ	試験運用期間	OK文 件数	OK文 検出件数	誤検出率	NG文 件数	NG文 検出件数	検出率	適合率
第1期	2018/7/23～2018/8/20	1207	146	0.121	41	24	0.585	0.141
第2期	20180831～20180914	1949	122	0.063	51	35	0.686	0.223



(a) LSTM 言語モデル　(b) BLSTM 言語モデル
(c) BLSTM OK/NG確率モデル Ⅰ　(d) BLSTM OK/NG確率モデル Ⅱ
(e) アンサンブルモデル

図 3: 文単位性能評価結果(ROC 曲線)