

日本語学習者向けの文法誤り検出機能付き誤用例文検索システム

新井 美桜 金子正弘 小町 守

首都大学東京

{arai-mio, kaneko-masahiro}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

近年、日本語を母語としない人々の日本語学習の需要が高まっている。日本語学習者は国内、海外共に年々増加しており、日本語学習を支援するための研究が注目され始めている。

学習者が作文する際有用なものが、用例検索システムである。用例検索システムとは多数の文を収録したコーパスから探したい語句を含む文を検索し、どのようにその語句が使われているか確認することによって学習者の作文を支援するシステムである。

既存の日本語学習者の作文用例検索システム [5][6] には、正用例のみまたは誤用例のみまたはその両方を検索するものがある。学習者は、作文を執筆する際、用例検索システムで使いたい語句がどのように使われているかを確認する。学習者の検索したい語句が正しければ、正用例のみを表示する通常の用例検索システムでも検索できる。しかし、学習者の想像している語句が間違っている場合、用例検索をすることができない。また、誤用例のみの検索ができたとしても、それに対する添削がなければ、それがどのように間違っていて、どのように直せば良いのかわからない。また、誤用例に正用例が付与されていたとしても学習者自身が誤りを認識していない場合、どの部分を直せばいいのかわからず、混乱を招く。

提案システム (Sakura) では、誤り検出機能を付与し、学習者が入力した文の誤りを自動で検出することで学習者が気づいていない誤りを認識でき、かつ正用例付き誤用例を表示することで学習者が解釈して訂正または訂正しないという選択ができる形の検索システムを構築した。

この検索システムは Web サイト¹にて公開している。

2 関連研究

2.1 日本語の用例検索システム

今日、様々な日本語学習者向けの用例検索システムが開発されている。しかし、未だ学習者が実用的に利用することは難しい部分が多く存在する。例えば、既

存の日本語学習者の作文用例検索システムには以下のようなものがある。

李ら [5] は日本語学習者の発話データを文字化した言語資料である『KY コーパス』に言語情報を付与した『タグ付き KY コーパス』に用例検索機能を追加し、公開した。『タグ付き KY コーパス』は検索したい語を入力するとコーパスから学習者の正用例、誤用例を検索できる。しかし、『KY コーパス』の誤用例は非常に数が少なく、頻繁に使用するであろう語句にも誤用例が存在しない場合が多くある。また誤用例に対応する正用例がないので、学習者が作文に一致する誤用箇所を見つけても、訂正の仕方がわからず、参考にすることが難しい。また、そもそも入力したクエリが誤用であると学習者自身が認識していないと誤用検索を行えないため、学習者が利用しづらい。

新井ら [6] は、誤用例の検索の有用性に着目し、相互添削型 SNS、Lang-8²を用いた日本語学習のための大規模誤用検索システムを構築した。Lang-8 データセットは大規模であるため、『タグ付き KY コーパス』のようなデータスパースネスの問題が解決できる。また、Lang-8 のデータセットは誤用例に日本語母語話者の書いた添削文が付与されており、このデータセットを誤用例検索に利用することで、学習者の文がどのように間違っていてどのように直せばよいのかの判断の参考にできる。しかし、正用検索と誤用検索を学習者が選択する形であり、学習者が自らその表現は誤っているという推測のもと検索しなければ適切な検索は行えないため、実用的に使うことが難しい。

本研究では、新井らと同様、Lang-8 データセットを用いることで、添削文付きの誤用例検索を可能にし、かつ、誤り検索機能を追加し、学習者が自分の誤りを認識していない場合の誤用検索も可能にした。

2.2 文法誤り検出

英語の文法誤り検出タスクではニューラルネットワーク、特に Bi-LSTM を用いた手法が盛んに研究されている。日本語の文法誤り検出の研究では、学習者の誤り種類を限定したもの、主に助詞に限定したものがほとんどである。[2][4] 関連する日本語の誤り訂正

¹<http://cl.sd.tmu.ac.jp/sakura/v3>

²<http://lang-8.com/>

モデル	precision	recall	F-value
比較 SMT モデル 提案システム	0.599 0.615	0.121 0.304	0.202 0.407

表 1: 日本語学習者の誤り検出精度

タスクでは、学習者の誤り種類を限定していないものには水本らの統計的機械翻訳を使用した研究がある。[7] 一方で、日本語ではニューラルネットワークを用いた文法誤り検出の研究は少ない。

本研究では、誤り種類を限定せず、ニューラルネットワークを用いての誤り検出器の構築に臨んだ。統計的機械翻訳は n-gram ベースであるため長距離の関係を考慮することができないが、ニューラルネットワークは入力履歴を状態変数として保持しておけるため長距離の関係を考慮することができる。

3 日本語学習者の誤り検出

3.1 誤り検出モデル

本研究では誤り検出を系列ラベリングタスクとして扱い、入力文の各単語に正誤ラベルを割り当てた。誤文と正文から動的計画法を用いて検出のためのラベル付けを行った。Rei ら [3] と同様に、誤り検出のための Bi-LSTM を構築する。システムは単語 $[w_1 \dots w_T]$ を入力として受け取り、各単語に対してラベルを予測する。全ての単語 w_t は、まず単語ベクトル e_t^w と文字ベクトル e_t^c に変換される。文字ベクトルは 1 文字を入力とする Bi-LSTM の先頭と文末の隠れ層を結合し作成する。Bi-LSTM に文字を入力する際は文字分散表現に変換される。これらのベクトルは、学習時に更新される。そして、 e_t^w と e_t^c を結合することで t 番目の入力ベクトル \tilde{x}_t を作成する。入力ベクトルは Bi-LSTM [1] を用いて以下のように隠れ層 h_t を計算する：

$$\vec{h}_t = LSTM(\tilde{x}_t, \vec{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = LSTM(\tilde{x}_t, \overleftarrow{h}_{t+1}) \quad (2)$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (3)$$

\vec{h}_t は順方向 LSTM、 \overleftarrow{h}_t は逆方向 LSTM、 h_t は両方向の隠れ層を結合したものである。次に、全結合層を使用して Bi-LSTM と出力層の次元数の差を緩和するための追加の隠れ層 d_t を計算する：

$$d_t = \tanh(W_d h_t) \quad (4)$$

W_d は重み行列である。そして、出力層と softmax 関数を用いて予測を行う：

$$P(y_t | w_1 \dots w_T) = \text{softmax}(W_o d_t) \quad (5)$$

W_o は出力重み行列であり、 y_t は予測ラベルである。



図 1: 誤用検索システム外観。1. 検索窓に入力された語句を検索する。2. 誤り検出を行い、誤っている部分は赤く表示される。3. 検出した単語を訂正箇所を含む誤り文と添削された正例文が対になって表示される。検索されたクエリの書いてあるボタンをクリックすると、全ての用例が表示される。

3.2 実験

本研究では水本ら [7] によって作成された学習者コーパス、Lang-8 Learner Corpora を使用した。Lang-8 Learner Corpora は学習者の作文とその添削、エッセイ ID、ユーザ ID、学習言語タグ、母語タグからなる。学習言語のタグが Japanese のエッセイ数は 185,991 である。

本研究では、Lang-8 Learner Corpora の日本語学習者が書いた添削前の誤用例と日本語母語話者が添削した正用例のペアのセット 140 万文対を抽出し、誤用の長さ 100 文字以下、レーベンシュタイン距離 7 文字以下の 72 万文対を用いた。コーパスに関しては、学習データを 720,000 文、開発データを 1,000 文、評価データを 1,000 文用いた。

ハイパーパラメータの設定は単語分散表現と単語 LSTM の次元数は 300、文字分散表現と文字 LSTM の次元数は 100 とした。adadelata (学習率: 1.0) でバッチサイズを 64 文として最適化した。

比較システムとして水本ら [7] の SMT システムを再実装した。単語-単語対応モデルを使用した。訂正結果から本誤り検出モデルと同様に動的計画法を用いて検出のためのラベル付けを行なった。

結果を表 1 に示す。Precision, Recall, F 値の全てで従来手法より良いことがわかる。

誤り文字列	正解文字列	旧 Sakura (正)	旧 Sakura (誤)	新 Sakura	誤り種類
おねさん	おねえさん	×	×	○	O
ニュージーランド	ニュージーランド	×	○	○	O
みなさん	みなさん	×	○	○	R
大体に	大体	×	×	○	R
疑問をして	疑問に思って	×	×	○	M
驚い	驚き	×	○	×	M
がもらえる	しかもらえない	×	×	×	A
稼ぐ	稼いだ	×	○	×	A
ちぎい	少ない	×	○	×	L
助けられる	できる	×	×	○	L
しましだ	いました	×	×	×	P
死んちゃう	死んじゃう	×	○	○	P
ハウス	家	×	○	○	X

表 2: 誤りタイプ別テストフレーズ例

誤りタイプ	頻度	正解率		
		旧 (正)	旧 (誤)	新
誤り全て	55	0.00	0.45	0.44
混同	19	0.00	0.37	0.32
語彙選択	16	0.00	0.38	0.19
脱落	8	0.00	0.75	0.75
誤形成	6	0.00	0.40	0.67
余剰	3	0.00	0.67	1.00
発音	2	0.00	0.50	0.50
その他	1	0.00	1.00	1.00
正例	55	0.90	0.15	0.85
全ての平均	110	0.45	0.30	0.65

表 3: システム別の各誤りタイプの頻度と正解率

4 誤用検索システム

本節では、本検索システムのアルゴリズム及び検索システムの評価実験について説明する。図 1 に本検索システムのインターフェースを示す。

4.1 用例検索アルゴリズム

前処理に関しては、新井ら [6] のものと同様に行う。用例のデータベースには 3.2 節で構築したデータを用い、学習者の書いた誤り文と日本語母語話者が添削した正例文を対し表示する。入力クエリは文または単語を想定している。

1. 入力クエリの誤りを検出。
2. 誤りが検出された場合、その単語を誤用例に含む文を検索。誤りが検出されなかった場合、クエリを正用例に含む文を検索。
3. 入力クエリのベクトルと検索された誤用例文のベクトルの \cos 類似度が高い順に、正用例と共

に表示。Word2Vec を使用し、入力クエリ内及び誤用例文内の単語ベクトルの平均を文ベクトルとした。

4.2 検索システムの評価実験

誤用例表示と誤り検出の有用性を確かめるため、比較システムを用いて以下の二つの方法で評価実験を行った。

内的評価 検索システムのコーパスに使用していない Lang-8 の学習者作文データから、フレーズをランダムに抽出し、比較システムと提案システムで検索した。実際に学習者が使用する時は、誤りと正用どちらも検索するため、誤りフレーズを 55 フレーズ、正しいフレーズを 55 フレーズ、合計 110 フレーズ抽出した。比較システムには新井ら [6] のもの (旧 Sakura) を使用した。この検索システムは入力したクエリを正用例で検索するか誤用例で検索するかをユーザが選択するものであるため、正用検索を選択した場合と誤用検索を選択した場合の両方を比較実験に使用した。提案システムはそのフレーズを含む文単位で、比較システムはフレーズ単位で検索した。

誤りフレーズの場合は、検索結果の上位 10 位以内に正解のフレーズを認識することができた数を、正しいフレーズの場合はそのフレーズを含む用例が上位 10 位以内に出てきた数を数えその正解率を測った。誤りフレーズの誤りタイプを混同 (A)、語彙選択 (L)、脱落 (O)、誤形成 (M)、余剰 (R)、発音 (P)、その他 (X) の 7 つに分類した。表 2 は誤りタイプ別のフレーズ例である。

各システムの誤りタイプ別の頻度と精度を表 3 に示す。

旧 Sakura (正) は比較システムの正用検索バージョンである。正しいフレーズでは最も高い正解率となっ

たが、誤りフレーズの検索クエリは誤り文字列であるため、正用例で検索をかけても一致するものがなく、適切な回答を得られることができなかつたため、正解率は0.00となった。

旧 Sakura (誤) は比較システムの誤用検索バージョンである。誤用例で検索をかけるため、正しいフレーズで検索した場合用例はほとんど出てこなかつた。誤りフレーズ検索では高い正解率を得られた。検索クエリに一致した誤り文字列を含む誤用例と、対になる正用例を検索できる。

新 Sakura が提案システムである。誤り全体の正解率は旧 Sakura (誤) よりも少し低いが、比較システムはユーザが任意で誤用検索と正用検索を切り替えなければならない。しかし、どのクエリが誤りでどのクエリが正しいかわからないため、その負担が大きい。提案システムはユーザがその切り替えをする必要なく、誤り検出機能により入力したクエリが正しいが誤りであるかをシステムが判断して適切な方を検索するため、誤りフレーズ、正しいフレーズの両方を含めた全体の平均はこのシステムが最も高い正解率を出した。

比較システムと比べて、提案システムは誤形成がよくとれた。このタイプは誤り箇所が明示的であるため、誤り検出機能の精度が良く、さらに類似度でソートしているため、入力クエリに似た文が上位に来るため、それを正しく直された文が表示できるからだと考える。また、脱落や余剰などの分かりやすい誤りは、問題なく誤り検出が機能し、高い正解率でとることができた。

一方で、語彙選択が検索しづらかつた。提案システムは検索結果を類似度でソートしているため、入力クエリと似たような語を使った文が上位にくる。入力クエリと全く異なる語彙が正解だった場合、その類似度は低くなるため、提案システムは適切な答えを出すことができない。また、誤り検出の精度が十分でないため、誤っている入力クエリが正しいと判断されることも多く、その場合自動的に正用検索になるため、正しく直された文が表示されないこともある。

外的評価 被験者実験では、新システムを以下の3つの設定で使用する。システム3が提案手法である。全てのシステムで3.2節で構築したデータを用い、比較箇所以外は全て新システムの設定を適用する。

- システム1: 誤り検出なし正用例のみ表示
- システム2: 誤り検出なし正用例と誤用例を表示
- システム3: 誤り検出あり正用例と誤用例を表示

システム1を使った作文とシステム2を使った作文の点数を比較して、誤用例と正用例の文対表示の有用性を確認する。システム2を使った作文とその作文を

システム3を使って直した後の点数を比較して、誤り検出機能の実用性を確かめる。

日本語を母語としない日本語学習者6人に日本語作文の問題を1人につき10問解答してもらった。全員、情報工学を専攻する大学院生で、母語は中国語である。日本語のレベルは、5人が日本語能力試験N1、1人がN2である。AテーマとBテーマに5問ずつわけ、Aテーマをシステム1を、Bテーマをシステム2を使って書いてもらった。その後、システム2で書いた作文をシステム3を通し校正してもらった。評価基準を統一するため、1問につき3文の制限を付けた。評価者に書いてもらった文章に減点法で点数をつけた。持ち点1人30点で、1つの文法誤りにつき1点引いていき、各システムごとに合計し、最終的な点数を計算した。点数の平均はシステム1が20.8、システム2が22.7、システム3が**24.2**と提案手法が最も高く、システム2と同じ点数も含め、6人中5人が提案手法で最も高い点数を出した。

5 おわりに

誤用例文検索の有用性に着目し、正用例付き誤用例文の表示を行った。さらに、誤りを自動で検出することで学習者が気づいていない誤りを認識でき、かつ用例があるので学習者が解釈して訂正または訂正しないという選択ができる。

内的評価と外的評価で共に提案手法が最も高いという結果が得られた。

謝辞

Lang-8のデータ使用に際して、快諾して下さった株式会社Lang-8社長喜洋洋氏に感謝申し上げます。

参考文献

- [1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, 9, 1997.
- [2] Hiromi Oyama and Yuji Matsumoto. Automatic error detection method for Japanese case particles in Japanese language learners. In *Corpus, ICT, and Language Education*, 2010.
- [3] Marek Rei and Helen Yannakoudakis. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of ACL*, 2016.
- [4] 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 榎井文人. 日本語学習者の作文における格助詞の誤り検出と訂正. 情報処理学会研究報告コンピュータと教育 (CE), 2003.
- [5] 李在鎬, 浅尾仁彦, 濱野寛子, 佐野香織, 井佐原均. タグ付き日本語学習者コーパスの開発. 言語処理学会第14回年次大会, 2008.
- [6] 新井美桜, 小平知範, 小町守. Lang-8を用いた日本語学習者向けの誤用例文検索システム. 言語処理学会第24回年次大会, 2018.
- [7] 水本智也, 小町守, 永田昌明, 松本裕治. 日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得. 人工知能学会論文誌. 2013.