間違えやすい漢字の誤用訂正システム

小川耀一朗（長岡技術科学大学学生）・山本和英（長岡技術科学大学）

1. 研究背景

自然言語処理の分野において、日本語学習者の作文の誤りを自動で訂正する研究が行われている。これは教師が学習者の作文を添削する作業の負担を軽減し、また学習者自身の自己学習を支援することが目的である。教育現場ではコンピュータを利用した語学教育（CALL: Computer Assisted Language Learning）が取り入れられるようになり、多くの場面で活用されることが期待できる。

多くの研究では学習者の誤りを限定している。特に、助詞の誤り検出・訂正のみに限定した研究の報告が最も多い(笠原 2013) (今村 2012)。助詞は日本語学習者にとって間違えやすいため助詞誤り訂正の需要は高いが、それ以外の誤りの訂正技術も不可欠である。学習者の誤りを限定しない手法もある(水木 2013)。この手法では学習者作文とその修正文が対応した学習者コーパスを用いて機械翻訳モデルを学習させ、誤り文を正しい文に「翻訳」させることで訂正を行なった。しかし、この手法ではコーパス中に存在しない誤りの種類は訂正することができない。例えば「他」を「池」に間違える、「与」を「写」に間違えるといった形の似ている漢字での誤用はコーパス中になかったため訂正することができなかった。しかし、日本語学習者、特に非漢字圏日本語学習者にとって漢字習得は困難である。非漢字圏学習者にとって馴染みのある表音文字であるアルファベットなどと大きく異なり、表意文字である漢字は種類が多く、形や音、意味といった大量の情報を覚える必要があるためである。従って形の似ている漢字で間違えてしまう可能性は十分にある。そこで、本研究では文中の形の似ている漢字での誤用を自動で訂正するシステムを開発した。

2. 形の似ている漢字の収集と言語モデルを用いた誤用訂正手法

2-1. 類似セット

非漢字圏日本語学習者の作文に見られた漢字の書き誤りを収集し、調査した研究がある（佐々木 2008）。この調査では、書き誤りの要因には「異なる部品の使用」と「1 画多いまたは少ない」が多く、全体的な形は把握しているものの、漢字の細部の間違いをしてしまう事例が多く見られた。そこで、形は似ているが異なる漢字の組を集めた類似セットを作成した。この類似セットには、「白」と「自」のような 1 画だけ違う漢字、「牛」と「午」のような 1 つの辺の長さが違う漢字、「複」と「復」のような部首が異なる漢字、「使」と「便」のような旁が異なる漢字などの組が含まれている。

類似セットは 608 組、1,942 字の漢字で構成される（以下、この 1,942 字を対象漢字と呼ぶ）。1 組には 2 つの漢字のペアだけでなく、3 つ以上の漢字が含まれる組もある。ま

た、2,136 字の常用漢字に該当する漢字は 980 字含まれている。

2-2. 訂正手法

文中の形の似ている漢字での誤用を訂正するシステムには、言語モデルと上記の類似セットを用いる。言語モデルとは、任意の単語列が与えられた時、ある単語がその後に来る確率を与えるモデルのことであり、音声認識や機械翻訳など自然言語処理の分野で広く用いられている。より自然な文には高い確率を与え、反対により不自然な文には低い確率を与える性質がある。

訂正の過程を説明する。まず、入力文の各文字のうち、対象漢字に該当する漢字を検索する。次に、文中の対象漢字の 1 つに着目し、その漢字と同じ組に属する漢字を類似セットから抽出する。そして、抽出した漢字で対象漢字を置換した文を生成し、候補文とする。元の文とそれぞれの候補文の言語モデル確率を算出し、もし候補文の確率が元の文より高いものがあれば、最も確率の高い候補文で元の文を置き換える。この操作を対象漢字の全てにおいて繰り返すことで訂正を行う。

もし文中に「自白」のように対象漢字が連続している場合は、1 文字ずつ置換を行なっても正しい判断ができない場合があるため、「白白」、「白自」、「自自」のように連続する対象漢字の全ての置換の組み合わせを候補文に含めるようにした。また、対象漢字のうち形態素解析により固有名詞と判定される漢字は訂正の対象から除いた。形態素解析には Mecab[1] と UniDic[2] 辞書を用いた。言語モデルには、約 1 億語を収録した現代日本語書き言葉均衡コーパス（BCCWJ）[3] を訓練させ、KenLM toolkit[4] を用いて 4-gram 言語モデルを構築した。

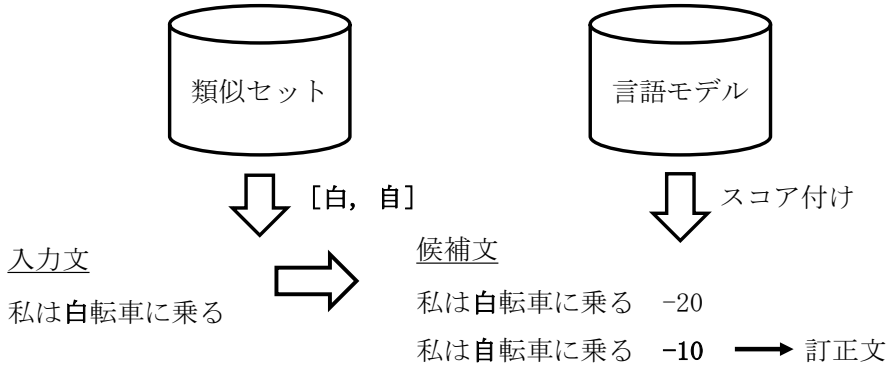

図 1　訂正システムの概略

2-3. 未知語フィルタ

漢字の誤用により、形態素解析を行うと辞書に存在しない「未知語」と判断されることがある。未知語と判断される漢字は周囲の単語と正しく組み合わされないため、多くの場合に誤用であると判断できる。また、文中に複数の誤用が含まれている場合、周囲の単語と連結して 1 つの未知語と判断されることがある。例えば「シドニー五論競泳伐表」は「輪」を「論」に、「代」を「伐」に誤用しており、混同して「五論競泳伐表」が 1 つの未知語と判断される。この場合、正用文と比べると分かち書きした時の単語分割数が少な

くなり、言語モデル確率を求める際の計算回数も少なくなるため、確率が高くなってしまう問題がある。そこで、言語モデル確率を計算する前の処理として未知語フィルタを挿入する。この未知語フィルタは、生成された候補文のうち、元の文と比べて未知語と判断される文字の数が増える文を除外する。これにより、正しくない候補文を選択することを防ぐ。

2-4. テストデータ

学習者コーパス「なたね」[(5)]及びオンライン日本語誤用辞典[(6)]から、漢字の形の類似が原因の誤用文を抽出し、47 箇所の誤用を含む 44 文を収集した。しかし、この学習者コーパスはシステムの汎用性を評価するには事例が少ないため、自動的に誤用を生成した擬似誤りコーパスを作成した。正しい日本語文中にある対象漢字を、類似セットの同じ組の漢字からランダムに 1 つ選択された漢字に置換することで、誤用を生成した。正しい日本語文には、毎日新聞コーパス[(7)]の一部を使用した。様々な誤用の事例を作るため、対象漢字それぞれに対して同じ操作を最大 10 文行ない、11,355 箇所の誤用を含む 10,677 文の擬似誤りコーパスを作成した。

2-5. 評価尺度

評価には、再現率、適合率及び F 値を用いた。再現率は誤用箇所のうち正しく訂正された箇所の割合を表し、適合率はシステムが訂正を行なったうち正しく訂正された箇所の割合を表す。F 値は再現率と適合率の調和平均である。

3. 訂正実験の結果と考察

学習者コーパスと擬似誤りコーパスの 2 つをテストデータとして実験を行なった。表 1 にコーパスごとの再現率、適合率及び F 値を示す。ベースラインとして頻度は候補の漢字の中から最も出現頻度の高い漢字を選択する方法である。出現頻度は事前に言語モデルの訓練に用いたコーパスから算出した。言語モデルを用いた手法と、さらに未知語フィルタを挿入した手法が提案手法である。頻度と比較して提案手法は高い精度を示している。

学習者コーパスでは、47 箇所の誤用に対して 44 箇所を正しく訂正し、再現率は 0.94 となった。またシステムは 50 箇所の訂正を行い、再現率は 0.88 となり、F 値は 0.91 となった。これより、ほとんどの誤用を正しく訂正できていることがわかる。また未知語フィルタは効果を発揮する事例がなかったため精度が変わらなかった。

擬似誤りコーパスは事例が少ない学習者コーパスと比べてかなり多くの事例を含んでいる。しかしながら、提案手法の精度を見ると高い値を示している。さらに言語モデルに未知語フィルタを挿入することで適合率が向上し、F 値は 0.910 となった。これより、提案手法は様々な誤りの事例にも高い精度で訂正を行うことができるため、汎用性が高いことを示した。未知語フィルタにより再現率がわずかに減少しているのは、正用箇所が未知語として使われている箇所があったからである。

表 1　2 つのコーパスでの誤り訂正結果

テストデータ	手法	再現率	適合率	F 値
学習者コーパス	頻度	0.40(12/47)	0.24(12/169)	0.30
	言語モデル	0.94(44/47)	0.88(44/50)	**0.91**
	未知語フィルタ+言語モデル	0.94(44/47)	0.88(44/50)	**0.91**
擬似誤りコーパス	頻度	0.476(3,844/11,355)	0.117(3.844/46,149)	0.188
	言語モデル	0.941(10,685/11,355)	0.856(10,685/12,478)	0.897
	未知語フィルタ+言語モデル	0.939(10,664/11,355)	0.883(10,664/12,076)	**0.910**

表 2　システムによる誤り訂正例

No.		文	備考
1	入力	日本人の矢ったものがここにある。	訂正成功
	出力	日本人の失ったものがここにある。	
	正解	日本人の失ったものがここにある。	
2	入力	その島は珍しい。	正しい箇所を間違って訂正してしまった
	出力	その鳥は珍しい。	
	正解	その島は珍しい。	
3	入力	ともだちのかぞくに合います。	誤用を訂正しなかった
	出力	ともだちのかぞくに合います。	
	正解	ともだちのかぞくに会います。	

表 2 にシステムの訂正例を示す。No. 2 は正しい箇所を間違って訂正してしまった例である。しかし出力文を見ると、文法的には間違っていないことがわかる。このような場合、言語モデル確率はどちらも高くなり、最終的には訓練コーパスでの出現頻度に依存する。言語モデルでは周辺の単語の情報しか活用しないため、正しい訂正をするには文の意味や前後の文の文脈を考慮する必要があるだろう。No. 3 は誤用を訂正しなかった例である。この文は学習者の作文であり、平仮名が多用されているが、平仮名の形態素解析は精度が高くなく、「ともだち/の/か/ぞ/くに/合い/ます」のように間違った単語分割がされてしまう。そのため正しい単語の情報を活用できず、訂正が正しく行われなかった。日本語学習者は漢字に代用して平仮名を多用する傾向があるため、平仮名の形態素解析技術の向上は誤り訂正タスクやその他の言語処理タスクにとって課題となる。

4. まとめ

本研究では、形の似ている漢字の誤用を訂正するシステムを開発し、高い精度で謝り訂正を行うことができた。実験には事例の少ない学習者作文コーパスと、多くの誤りの事例を含んでいる擬似誤りコーパスの 2 つを使用した。どちらのコーパスにおいても高い性能で訂正を行うことができることから、提案システムは様々な誤りの事例に対応することができることを示した。

本研究の成果物については、収集した形の似ている漢字のリストを含め、デモシステムを公開している[8]。

謝辞

本研究は、平成 27～31 年科学研究費補助金基盤(B)課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」の助成を受けています。

注

(1) http://taku910.github.io/mecab/
(2) http://pj.ninjal.ac.jp/corpus_center/unidic/
(3) http://pj.ninjal.ac.jp/corpus_center/bccwj/
(4) https://kheafield.com/code/kenlm/
(5) https://hinoki-project.org/natane/
(6) http://cblle.tufs.ac.jp/llc/ja_wrong/
(7) http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
(8) http://www.jnlp.org/SNOW/S14

参考文献

(1) 笠原誠司・藤野拓也・小町守・永田昌明・松本裕治 (2012)「日本語学習者の誤り傾向を反映した格助詞訂正」『言語処理学会第 18 回年次大会』pp.14-17, 言語処理学会.
(2) 今村賢治・齋藤邦子・貞光九月・西川仁 (2012)「小規模誤りデータからの日本語学習者作文の助詞誤り訂正」『自然言語処理』19 (5), pp.381-400, 言語処理学会.
(3) 水本智也・小町守・永田昌明・松本裕治 (2013)「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』28 (4), pp.420-432, 人工知能学会.
(4) 佐々木良造 (2008)「マレー人日本語学習者の作文にみられた漢字の書き誤り」『世界の日本語教育』18, pp201-213, 国際交流基金.