日本語学習支援のためのコーパス構築システム

高橋 哲朗
Fujitsu Laboratories of America, Inc.
(現 富士通研究所)
takahashi.tet@jp.fujitsu.com

相川 孝子
Massachusetts Institute of Technology
taikawa@mit.edu

1 はじめに

テクノロジの活用が各領域で進む中で，言語教育においても MOOCs(Massive Open Online Courses) を始め様々なオンラインの個人学習環境が提供されている．MOOCs を例に挙げてその学習の形態を見てみると，ビデオによる講義が中心であり，ビデオを観た後にその内容に沿って出題された問題を学習者が回答するという形式が主である．

これに対して実際の教室で行なわれている日本語の授業では，教師は講義や課題の出題だけではなく，学習者の間違いに対して適切なフィードバックをインタラクティブに行なっている．オンライン学習にもそのインタラクティブ性を追加できれば，より効果的な学習が実現できると考えらえる．

本研究ではインタラクティブな日本語学習支援システムを開発することをゴールとして設定し，そのプロトタイプシステムと，システムの用いる知識となる誤用・訂正パタンを獲得するためのコーパス構築システムを開発した．

2 日本語学習支援システム

開発中の日本語学習支援システムの外観を図 1 に示す．学習者がテキストを入力すると，システムは全てのキー入力を監視し，誤りを検出し次第すぐに誤り部分を指摘し，誤りの理由と訂正案を提示する．

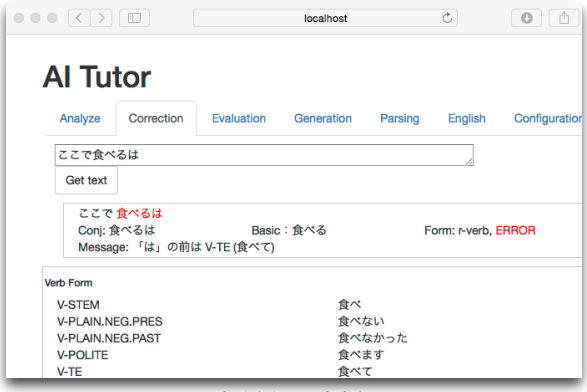

図 1: 日本語学習支援システム

図 1 の例では，学習者が最後の「は」を入力した時点で「食べるは」という表現が誤りであるとシステムが判定しメッセージを出している．この誤り判定を実現するために，システムは図 2 に示すパタンを用いている．このパタンは，一段活用型で活用形が基本形である動詞の後ろに「は」が続くのは誤りであり，その場合 “V-TE(動詞の TE-form)”[1] でなければならない旨のコメントを出すことを指示した知識である．このように抽象化した形態素の連接という形でパタンを人手により記述している．

```
[*; 一段; 基本形][は;*;*],「は」の前は V-TE
```

図 2: 誤り検出・訂正パタン

図 1 ような学習環境を提供することにより，学習者は誤りに対してリアルタイムにフィードバックを得ることができ，より効果的な個人学習を実現できると考える．このシステムは図 2 のような知識を用意することで学習者の誤りを自動的に判定できるようになるが，ここで課題となるのはこの知識をどのように構築するかである．誤りの種類は非常に多様であるため，この知識の構築は簡単ではない．この課題に対して本研究では日本語教師たちに知識を追加してもらう枠組みを考案した．我々はこの手法を言語教師という専門家によるクラウドソーシングと位置付け “Teacher sourcing” と呼ぶ．

3 Teacher sourcing による誤用コーパスの作成

北米において日本語を教えている 10 名の教師に依頼し，以下の 3.1 節，3.2 節で述べるタスクを実施した．

3.1 誤用例文の作成

各教師に一人あたり 200 文の日本語誤用例文を作成していただいた．誤用例文の作成にあたっては普段の授業や課題の中で目にする誤りを含む文を作文するよう依頼した．

10 名の日本語教師が誤用を含む例文 200 文をそれぞれ作成したが，この例文の中には重複があったため最終的に得られた誤用例文の異り数は 1,988 文であった．

[1] IPA 辞書では活用形が未然形

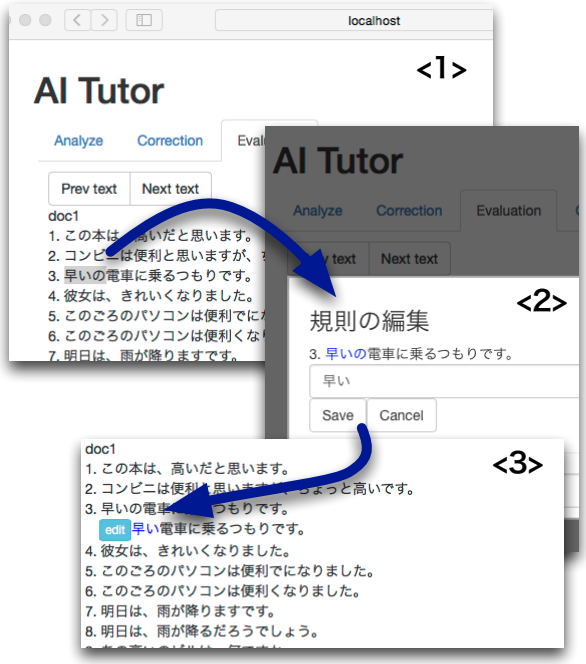

図 3: システム上での訂正

3.2 誤用例文に対する訂正

訂正タスクのために 図 3 に示す Web インターフェースを持つシステムを筆者らが用意し，各教師は Web ブラウザ上で訂正タスクを行なった．このシステム上で教師はまず，表示された誤用例文の中で誤っている個所の範囲を指定する＜1＞．すると規則の編集のための画面がポップアップ表示されるので，そこに訂正後の文字列を入力し保存する＜2＞と，その訂正パタンを用いて訂正された結果が元の画面に表示される＜3＞．

10 名の日本語教師によって作成された 1,988 文の誤用例文からランダムに 400 文を抽出し，各教師の訂正対象とした．

3.3 収集した訂正パタン

今回収集したコーパスには以下の情報が記録されている．

- 訂正前の文 (e.g. ピザを**食べるでしたか**)
- 指定した訂正範囲 (e.g. 4 文字目〜10 文字目)
- 訂正後の表現 (e.g. **食べましたか**)

10 人の教師が 400 件ずつ訂正を行なったが，1 人が複数個所を訂正したケースや，複数種類の訂正をしたケースがあったため，一人あたりの件数は 400 件以上となり，作成された訂正パタンの総数は 4,471 件であった．これらの訂正パタンを元に以下の分析を行なった．

表 1: 誤り範囲指定の一致

	2 人	3 人	合計
完全一致	617	19	636 (32%)
部分一致	1,286	5	1,291 (65%)
不一致	53	0	53 (3%)
合計	1,956	24	1,980 (100%)

3.3.1 誤り範囲指定の一致の分析

今回対象とした 1,988 文の誤用例文のうち，2 人または 3 人によって訂正パタンの作成が行なわれた事例の数は 1,980 件であった．この中で，教師が誤りとして指定した範囲が一致した数を表 1 に示す．

3 人によって訂正パタンが作成された場合は，そのうちのいずれか 2 人の指定範囲が一致していれば一致として集計した．部分一致には以下の (1), (2) のような訂正パタンが含まれる．この例では “**降ります**” が “**降りますらしい**” に部分一致している．

(1) 明日は雨が “**降ります**” らしいです。→
明日は雨が “**降る**” らしいです。

(2) 明日は雨が “**降りますらしい**” です。→
明日は雨が “**降るらしい**” です。

完全一致の二倍以上の訂正パタンが部分一致であり，教師による範囲指定の揺れの大きさを示している．NAIST 誤用コーパスを開発した大山ら [2] も誤用タグの範囲を決めることの難しさを議論している．本研究では今後複数人の教師によるタグ付けを通してその範囲の多様性を調査していきたい．

範囲指定が一致した場合に訂正文字列が一致した割合は 567/735(77%) であった．訂正文字列が異なる例を見てみると (3) や (4) のようにどちらの訂正文字列も正しいと考えられた．

(3) 昨日は “**雨だでしょう**” → 雨だった／雨でした

(4) 町は “**静かくないです**” →
静かじゃありません／静かではありません

3.3.2 再利用可能性の分析

今回作成された訂正パタンの異なり数は 3,766 件であったが，このうち 73 件の訂正パタンが事例をまたいで一致した．つまりこの数の訂正パタンが再利用可能であったと言える．今後大規模に Teacher sourcing が行なわれることにより再利用可能な訂正パタンの数は増えていくことが期待できるが，それだけではなく規則の一般化も必要である．

たとえば今回作成された訂正パタンの中に (5)〜 (7) があったが，これらは (8) のような訂正パタンに抽象化が可能である．

(5) げんきかった→げんきだった

(6) きらいかった→きらいだった

(7) べんりかった→べんりだった

(8) (名詞-形容動詞語幹) かった→
(名詞-形容動詞語幹) だった

一方で (9) のように形態素解析できない事例も多く存在するため，形態素を用いた訂正パタンには限界があり，表層単位でのパタンの獲得も重要であると言える．

(9) すわてもいいです

4 議論

4.1 日本語教師の負荷とならないタスク設計

日本語教師は普段の授業で学生に課題を出し，その内容を確認しフィードバックをしている．図 4 は実際に学生によって提出された課題を日本語教師である第二著者が訂正した結果である．ここで行なっているタスクは，まさに 図 3 で示した誤用例文に対する訂正のタスクと同様であるため，図 3 に示すシステムを学生の提出した課題を評価するツールとしても用いることにより，新たな時間をかけずにコーパスを作成できると期待している．

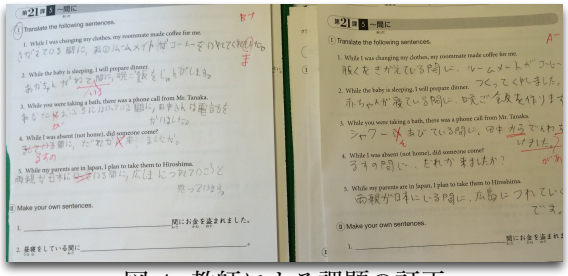

図 4: 教師による課題の訂正

4.2 システムによる誤り判定・訂正の自動化

今回の訂正タスクでは，複数の教師間での指定する範囲の一致度などを調査するために，他の誤用例文で作成された訂正パタンは再利用されないようにした．しかし，実際に本システムを用いて学生の課題を評価する際には，入力された誤り訂正パタンはシステムにより再利用可能な形で蓄積されるため，即座に他の文にも適用され自動的に訂正案として出力できる．つまり，教師がこのシステムを使えば使うほどより多様な誤りに対応できるようになる．

図 5 の例では，他の誤用例文「このサービスは便利と思います。」で作成された訂正パタンが 2 番目の例文に適用されている例を示している．教師は再度訂正を行なう必要はなく，システムの提示した訂正を確認

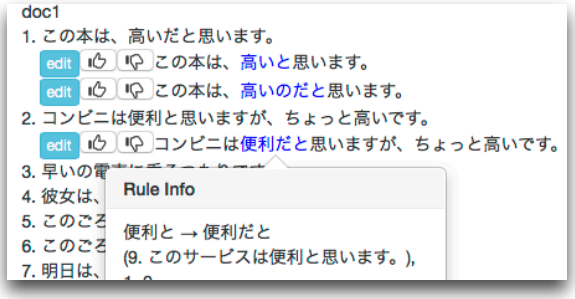

図 5: 候補選択および規則の再利用の例

するだけで良い．これにより教師の負荷を軽減させ，まだ訂正されていない種類の誤り訂正に注力できるようになる．実際の授業においては同一の課題を全学生に課すことが多く，その際同じ課題に対して学生たちは共通した誤りを起こしやすい．このような場合には特にシステムによる自動訂正が教師の時間を節約するために有効である．

4.3 教師のフィードバックによる知識の学習

複数の訂正パタンが適用可能な場合は図 5 の 1 番目の文のように，複数の訂正候補が出力される．ここで教師は thumbs up ボタン ( ) を押すことにより，システムが生成した訂正案を採用できる．また正しくない訂正案に対しては thumbs down ボタン ( ) を押すことにより，訂正案が正しくない旨をシステムにフィードバックできる．

これら両方のフィードバックは訂正パタンを洗練するための有用な情報となる．たとえば訂正個所が文脈に依存する場合には，教師によって押された “thumbs down” ボタンによりその知識が特定の文脈では使えないことをシステムが知ることができる．このようなフィードバックを多く集めることにより，訂正が可能となる条件を “thumbs up” や “thums down” の付けられた事例から学習することが可能となる．

4.4 教師間の知識共有

本システムでは Teacher sourcing により他の教師によって作成された知識によって訂正された候補も提示される．これにより誤りの範囲の指定方法や訂正方法に関する多様な知識の共有が実現でき，教師たちの指導内容の幅を広げることができる．そして，集約された知識や用例を総合的に用いることができるようになるため，言語教育そのものを洗練していくことが期待できる．

4.5 今後の発展

今後，教師による利用をより進めてコーパスを拡大していくことを計画している．また 3.3.2 節で述べたように，今回収集した誤り訂正のパタンには，教師による大きな揺れがあることが分かったので，これらの揺れを吸収し，多様な文脈でも適用可能な訂正パタンとしていくために，適切な範囲でのパタンの抽出とパタンの一般化を行ないたい．

図 3 で紹介した Teacher sourcing のインターフェースでは，文字列または形態素単位の連接で表記できる誤りしか扱うことはできない．将来的には呼応などの複雑な構文的パタンや，より意味的な処理を必要とする文脈解析などの領域においてもシステムによる誤り判定や訂正ができることが望ましい．しかし現時点ではこれらの知識を記述し活用することが困難であるため，本研究のスコープ外とし，これらの問題の教示は実際の教師に委ねたい．

5 関連研究

これまでに複数の研究機関により日本語学習者の誤りの用例やメタ情報が付与されたコーパスが作成されている [4, 5, 7, 3, 6]．これらの先行研究とは異なり，本研究では誤り判定および訂正の知識の収集のために，教師によるクラウドソーシングである Teacher sourcing を用いた．また単に言語データとしてのコーパスを作成するだけではなく，蓄積された知識をすぐにシステムに再利用させることにより，不足している知識を効率良く追加できるサイクルを実現している点が，従来の研究とは異なる．

Mizumoto ら [1] は 言語学習 SNS Lang-8 のログデータである誤り文と訂正文から誤用事例を抽出し，エラー判定や訂正に用いており，本研究とはこの点で共通している．しかし Mizumoto らが統計的機械翻訳の手法によりエラー訂正を行なったのに対し，本研究では訂正パタンの抽出に留めた．その理由の一つは，本研究の対象が言語教育であるため，教師が教えている方法で誤りを説明できなければならないと考えたからである．また教師間の知識の共有や，教師のトレーニングといった効果を期待していることも理由の一つである．

6 おわりに

テクノロジを活用した教育システムはこれまでに多数提案されてきているが，そこで用いられている技術は動画配信，ビデオチャットによるコミュニケーション，SNS を用いた教材の配布など，教育外の分野で開発された技術が主であり，それらの技術を受身的に教育分野に適用した事例であると言える．これに対し，本研究で目指しているのは現場の教師たちが暗黙知として持っている知識を Teacher sourcing の手法を用いることにより表出させ，再利用可能な形式知にし，その知識を語学教育に還元するという試みである．またこれは上述した動画配信などの技術の利用とは本質的に異なり，言語教師と自然言語処理の研究者との密接な協力がなければできないことでもある．

本研究を進める上では言語教師と自然言語処理の研究者の協力が必要になり，従来同じ「言語」という対象物を扱っている分野でありながら交流の少ない 2 つの分野の距離を狭めてくれると期待できる．

謝辞

本研究における Teacher sourcing は Japan Foundation, Los Angels の助成を受けて行なったものであり，ここに感謝を記したい．

参考文献

[1] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 147–155, 2011.

[2] 大山浩美, 小町守, 藤野拓也, 松本裕治. 日本語学習者の作文におけるエラータイプの自動分類へ向けて. 第三回コーパス日本語ワークショップ, 2013.

[3] 大山浩美, 小町守, 松本裕治. 日本語学習者の作文における誤用タグつきコーパスの構築について—NAIST 誤用コーパスの開発—. 第一回テキストアノテーションワークショップ, 2012.

[4] 寺村秀夫. 外国人学習者の日本語誤用例集. 大阪大学; データベース版、国立国語研究所、2011 年, 1990.

[5] 大曽美恵子, 杉浦正利, 市川保子, 奥村学, 小森早江子, 白井英俊, 滝沢直宏, 外池俊幸. 日本語学習者の作文 コーパス:電子化による共有資源化. 言語処理学会第 3 回年次大会論文集, p. 131145, 1997.

[6] 通子望月. 日本語教育における学習者コーパスの構築とicleaj. 関西大学外国語学部紀要, No. 7, pp. 111–119, oct 2012.

[7] 李在鎬, 林情, 宮岡弥生, 柴崎秀子. 言語処理の技術を利用したタグ付き日本語学習者コーパスの構築. 2012 年度日本語教育学会春季大会予稿集, 2012.