識別的系列変換を用いた日本語助詞誤りの訂正

今村 賢治　　齋藤 邦子　　貞光 九月　　西川 仁
日本電信電話株式会社，NTTサイバースペース研究所
{imamura.kenji, saito.kuniko, sadamitsu.kugatsu, nishikawa.hitoshi}@lab.ntt.co.jp

1　はじめに

日本語学習者の作文の誤り訂正は，従来から教育の一環として研究されてきたが，近年はビジネス上の必要性も生じてきている．たとえば，オフショア開発（システム開発の外国への外部発注）では，中国，インドなどへの発注が増加している．外国に発注する場合，日本との意思疎通は英語または日本語で行われるが，日本語学習者の多い中国北部では，日本語が使われることも多い．しかし，中国語を母語とするものにとって日本語は外国語であり，メールなどの作文には誤りを含み，意思疎通に問題となるため，それらを自動検出・訂正する技術が望まれている (大木他, 2011)．

本稿では，中国語母語話者の日本語作文の誤り自動訂正について述べる．外国人にとって，助詞はもっとも誤りやすい語であるため，本稿では助詞の用法を訂正対象とする．

2　日本語学習者の誤り傾向

まず，実際に外国人がどのような日本語書き誤りをしてしまうのか，誤り例を収集した．

被験者となる中国語母語話者は日本語の学習歴があり，日本の技術系大学に在籍する，もしくは卒業した背景をもつ37名である．日本滞在歴は半年から6年程度である．各被験者に技術系文書 (Linuxマニュアル等80文) の英文と24個の図 (のべ104課題) を提示し，キーボード入力による日本語作文を実施した（これを学習者作文と呼ぶ）．最終的には2,770文の学習者作文データを収集し，各作文を日本語母語話者が推敲した（以下，単に修正文と呼ぶ）．

2.1　誤りの分類と出現分布

誤り傾向の分析にあたり，まずは大分類として，文法誤り，語彙誤り，表記誤りの3種類を設定し，さらに小分類を設定した（表1）．

収集した2,770文の分析を実施したところ，訂正が可能であったものは2,171文であった．訂正が出来なかったものは，全く誤りがない日本語文559文，および文として成り立っておらず，手の施しようのない日本語文40文である．これ以降の分析は，訂正が可能であった2,171文に対して行った．

まず，誤り訂正の発生箇所は4,916箇所であり，1文あたり平均2.26箇所であった．また各誤りの種別について，誤り大分類での出現分布をみると，文法誤りが54%と最も多く，続いて語彙誤り28%，表記誤りが16%であった．これ以外は複数の誤りが混在する複合型誤りである．さらに小分類での出現分布をみると，最も多く発生していたのは助詞誤り33%，続いてカタカナ語誤り11%，単語選択（類義語）の誤り10%であった．対象文書が異なっているにも関わらず，文法誤りが約半数，助詞誤りが全体の1/3程度という結果は，大木他 (2011) の分析とほぼ一致している．

2.2　誤り傾向

今回の誤り傾向であるが，助詞誤りおよびカタカナ誤りは中国人に限らず広く外国人に共通して出現するものであると推測される．助詞は日本語特有の文法であり，多くの非日本語母語話者にとっては習得が難しいものである．そのため，中国人に限らず外国人の学習者作文の誤りに対する訂正対象を助詞とすることは，発生率から考えても効果的である．また，カタカナ語については英語の発音として覚えた単語をそのままカタカナで表現する際に，本来の日本語カタカナとは異なる文字へ変わってしまうことがよくあると考えられる．

助詞の種類によって誤り発生のしやすさは異なっているはずであり，全ての助詞が一律に誤りとはならない．今回の作文データにおける助詞誤りについて，さらに詳細に内訳を分析をしたところ，まず，誤りタイプとしては置換誤りが74%，助詞の抜けが17%，余分な助詞の出現が9%であった．特に置換誤りの発生が高い．また余分な助詞の出現が9%と非常に低く，訂正のために助詞の削除操作が必要となるケースは少ないことがわかる．個別の助詞誤り発生回数上位10件は表2のとおりである．

表 1: 誤りの分類と誤り例

大分類	小分類	誤り例
文法誤り	助詞，活用，接続詞，指示詞，疑問詞，語順，態，時制	[助詞] 質問**を**対応する [活用] ブック**を開きて**ください [接続詞] 20MB を超える**だから**アップロードします [指示詞] **その**以下のサイズに設定
語彙誤り	同音異義語，単語選択（類義語），母語の混用	[同音異義語] メモリ**内臓** [類義語] **快速**に処理します
表記誤り	カタカナ語，促音長音濁音，誤字脱字	[カタカナ語] アイコン**を**クリークする [促音長音濁音] 質問が**あたら**お願いします [誤字脱字] 私**立ち**でやります

表 2: 頻出した助詞誤り

誤り	正解	訂正タイプ	頻度
は/係助詞	が/格助詞	置換	117
を/格助詞	が/格助詞	置換	87
	の/連体化	挿入	70
を/格助詞	に/格助詞	置換	69
を/格助詞	が/格助詞	置換	66
	を/格助詞	挿入	65
が/格助詞	は/係助詞	置換	65
の/連体化		削除	61
は/係助詞	を/格助詞	置換	54
	に/格助詞	挿入	49

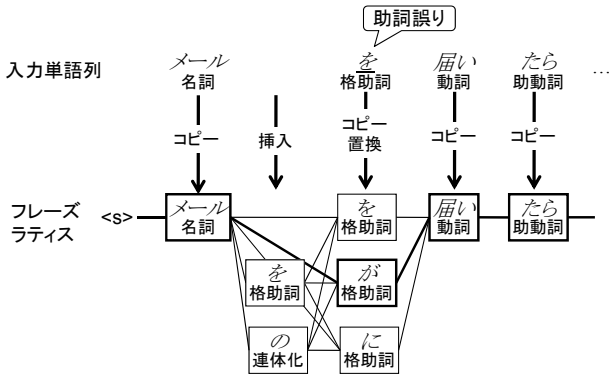

図 1: フレーズラティスの例

このうち,「は」と「が」の置換については，日本人でもこの使い分けは難しい場合もあり，必ずしも誤りとは言いきれないものも含まれる.「の」の助詞抜けとしては,「2 つファイル」のように，数量表現に後続する名詞の直前の「の」が欠けている誤りが良く見られた．また，余分な助詞「の」としては,「やったの人」「小さいの絵」など，動詞や形容詞に後続して「の」が余分に存在している誤りが多い．

以上の分析から，誤りの出現頻度の高い，助詞誤りを訂正対象とする．また，助詞の置換，挿入，削除が現れていることから，原文（入力文）を置換，挿入，削除操作することにより，誤り訂正を行う．

3 識別的系列変換

本節では，誤り訂正方式について述べる．本稿の誤り訂正は，学習者作文および修正文をあらかじめ形態素解析し，単語列から単語列へ変換することで行う．これは語順変更のない統計翻訳と同等であるため，同等な機能を持つ形態素変換器 (Imamura et al., 2011) をベースにする．以下，形態素変換器を利用した誤り訂正方法について説明する．

3.1 基本方式

形態素変換を使用した場合，以下の手順で入力文の誤りを訂正する．

- まず，入力単語列でフレーズテーブルを検索し，入力側にマッチするフレーズを得る．フレーズテーブルは，助詞誤りとその訂正候補を格納したもので，表 2 をテーブル化したものである[1]．フレーズテーブルと照合することにより，すべての訂正候補が得られる．また，無修正の場合を考慮し，入力単語を出力単語にコピーしたフレーズを作成し，両者をまとめてラティス構造にパックする（図 1）．これをフレーズラティスと呼ぶ．
- フレーズラティスから，条件付き確率場 (CRF (Lafferty et al., 2001)) に基づき，最尤フレーズ列を探索する．本稿の誤り訂正では語順の変更を行わないため，探索には Viterbi アルゴリズムを用いる．
- 学習時には，学習者作文と修正文に対して，DP マッチによる単語アライメントを行い，正解のフレーズ列を作成する．この正解から，助詞誤りだけを取得してフレーズテーブルを作成するほか，正解を教師データとして CRF を学習する．[2]

3.2 挿入・削除操作

一般的に句に基づく翻訳器は置換操作のみで翻訳を行うが，本稿で実施する誤り訂正は，助詞の置換操作のほかに，挿入，削除操作も対象となる．挿入操作は，

[1] 表 2 はフレーズテーブルの一部で，実際にはすべての助詞誤りを対象とした．

[2] 本稿では，CRF 学習のための最適化プログラムとして，岡崎の libLBFGS を用いた．
http://www.chokkan.org/software/liblbfgs/

空単語からある単語への置換，削除操作は，ある単語から空単語への置換とみなせるため，両者も基本的には置換操作と同等に扱い，モデルの学習・適用を行う．

しかし，挿入操作は，全単語間に挿入される可能性があるため，ラティス構築時にサイズが爆発するなど，非常に計算コストの高い操作である．挿入箇所をある程度絞ることが望ましいため，本稿では，名詞直後に後続する助詞のみ，挿入を許可するという制約をかける．

3.3 素性

本手法では 2 種類の素性を用いる．一つは翻訳モデルに相当する入力と出力のフレーズ対応度を測るためのマッピング素性，もう一つは言語モデルに相当する出力単語列の日本語としてのもっともらしさを測るためのリンク素性である．

固有表現抽出など，識別モデルを用いるタスクでは，タグを付与すべき単語のほかに，その周辺単語を素性として用いる場合が多く，今回も同様な考え方をする．具体的には，当該フレーズの入力側前後 2 単語をウィンドウとして，1〜3-gram と当該フレーズの出力単語の対を，二値のマッピング素性として使用する．

誤り訂正タスクにおいては，「正しい日本語」を出力する必要があるため，リンク素性は重要であると考えられる．幸いにも，識別モデルを用いる本稿の方式は，相互に依存する素性を混在させることができるため，以下の 2 種類のリンク素性を併用する．

- n-gram 素性: 出力単語の 1〜3-gram を二値素性として使用する．訓練コーパスからしか獲得できない．個々の n-gram の素性重みは，他の素性との兼ね合いを考慮しながら最適化されるため，きめ細かい最適化ができ，訓練コーパスにおける精度は高い．
- 言語モデル確率: 出力単語列の n-gram 確率（実際には 3-gram 確率）の対数値を実数素性として使用する．素性重みは 1 つしか付与されないが，言語モデルは「正しい日本語文」があれば学習できるため，訓練コーパスに限らず，大量の文から構築できる．

識別学習における二値素性と実数素性の混在は，半教師あり学習における補助モデル (Suzuki and Isozaki, 2008) と同じ考え方であり，訓練コーパス上での精度を保ちながら，未知テキストに対して頑健な訂正が行えるという利点がある．

4 誤り訂正実験

4.1 実験設定

コーパス 実験に使用したコーパスは，2 章で述べた 2,770 文 (104 課題) である．ここから助詞誤りのみを残し，それ以外の部分は日本語修正文の単語を埋め込んだ文を作成，コーパスとした．つまり，実験に使用した文ペアは，助詞誤りのみを含んだものである．誤り総数は，1,153 箇所となった．誤り助詞と訂正助詞を対にした異なり数は，154 種類（置換修正 110 種類, 挿入 21 種類, 削除 23 種類）である．なお，実験に使用したすべての文は，MeCab[3]によって形態素解析し，その表記と品詞を単語情報とした．

言語モデル 言語モデルは，Wikipedia のコンピュータ関連記事と，CentOS 5 の日本語マニュアルから，のべ 527,151 文を取得し，SRILM(Stolcke et al., 2011) でトライグラムを学習して使用した．

評価法 評価は，コーパスを課題単位に分割し，5 分割交差検定で行った．正解の単語列とシステム出力の単語列を比較し，誤り訂正の再現率，適合率を算出した．最終的な評価基準は F 値を利用した．

4.2 実験結果

実験結果を表 3 に示す．なお，表 3 で比較した手法は，以下の 3 種である．

- **提案手法:** リンク素性に n-gram 素性，言語モデル確率を併用した場合．
- **n-gram 素性のみ:** リンク素性に n-gram 素性のみを用い，言語モデル確率を使用しない場合．
- **言語モデル確率のみ:** リンク素性に言語モデル確率のみを用い，n-gram 素性を使用しない場合．

まず，トータルについて，使用したリンク素性を比較すると，F 値に関しては，この実験では提案手法（両者併用），言語モデル確率のみ，n-gram 素性のみの順で精度がよい．n-gram 素性と言語モデル確率は，モデル構築に用いるコーパスサイズに依存すると考えられるので，どちらが効果的かは一概には言えないが，両者を併用すると効果は大きい．特に再現率が大幅に向上（10.1%, 12.7%→ 21.6%）したことにより，F 値を向上させたことがわかる．適合率については，リンク素性を変えても変化は小さい．この傾向は置換修正，挿入修正に分けたときも同様だが，削除修正に関しては，言語モデルを使用しない（n-gram 素性のみ）ときに適合率が最高になった．言語モデル確率は総単語

[3] http://mecab.sourceforge.net/

表 3: 誤り訂正結果
トータルの適合率 (Prec.)，再現率 (Rec.) と F 値，および正解における置換，挿入，削除訂正別の精度．カッコ内の数値は，正解における訂正数を表す．

リンク素性	トータル (1,153)			置換 (762)		挿入 (264)		削除 (127)	
	Prec.	Rec.	F	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
提案手法	**52.4%** (249/475)	**21.6%** (249/1153)	**0.306**	**54.8%**	**23.9%**	**58.8%**	**15.2%**	36.0%	**21.6%**
n-gram 素性のみ	50.0% (117/234)	10.1% (117/1153)	0.169	50.3%	12.7%	42.9%	4.5%	**61.5%**	6.3%
言語モデル確率のみ	47.6% (146/307)	12.7% (146/1153)	0.200	51.3%	15.5%	43.8%	2.7%	34.4%	16.5%

数が少ない文に高確率を与えるため，過剰に削除される傾向があることを示している．

約 52%の適合率は，48%程度の修正箇所を再修正しないと正しい文にならないという意味で，実用上は決して高いとは言えない．助詞の用法には，意味的・文法的に明らかな誤用と，許容可能なものがあるため，人手評価を行った．

何らかの修正操作を出力したが，正解と異なった部分 226 箇所に関して，1 名の評価者によって主観評価した．なお，そのうち 186 箇所は，正解では無修正だった部分を過剰に修正したものである．評価観点は，システム修正を許容可能か（正解と比較して，意味的・文法的に異なっていないか）である．結果，226 箇所のうち 135 箇所は許容可能であった．つまり，許容可能という観点での適合率は，(249+135)/475=80.8%である．

なお，マッピング素性に関する補助モデルも，方式的には導入可能であるが，学習者作文・修正文ペアを大規模に集める必要がある．

5 関連研究

日本語学習者の助詞誤り検出・訂正は，従来より研究されており，大木他 (2011) は，形態素・構文解析済みの入力文（誤りを含む）に対して，周辺の形態素や係り先を素性として，SVM で助詞の誤用検出する方法を提案している．ここでは，助詞の欠落も対象としている．検出を行うのみで修正までは行わない．

Suzuki and Toutanova (2006) は，最大エントロピー法による分類器を用いて，助詞（主に格助詞）が欠落した文からの復元を行っている．この入力文は形態素・構文解析済みであり，基本的に誤り箇所が既に分かっているとき，挿入操作だけで修正を行う．識別モデルを使用しているため，言語モデル確率も素性として利用可能であるが，実際に適用まで行っていない．

助詞誤りに限定せず，すべての誤りを対象とした自動訂正には，統計翻訳を用いた Mizumoto et al. (2011) の方法がある．ここでは，対訳文に相当する学習者作文と日本人による修正文のペアを大量に SNS から収集し，句に基づく統計翻訳の仕組みを利用して訂正を行う．誤りを含む入力の形態素解析は行わず，文字単位で翻訳を行う．置換操作のみで修正するため，抜け落ちた助詞の復元のような挿入操作は，前後の文字と結合させた上で置換操作しなければならず，かなり大規模な学習者作文・修正文コーパスが大量に必要である．

今回用いた方法は，基本的には統計翻訳と同様な手法である．ただし，MOSES などの一般的な句に基づく統計翻訳が複数の生成モデルのスコアを対数線形結合しているのに対して，本稿では識別モデルの二値素性を主に用い，言語モデル確率を補助モデルとして使用する．また，置換操作のみでなく，挿入・削除操作も用いて訂正を行うのが特徴である．

6 おわりに

本稿では，中国語母語話者の日本語作文の誤り傾向分析と，助詞誤りに限定した誤り訂正法について述べた．すべての誤りを対象にしたわけではないが，誤りの 21.6%を 80.8%の適合率で許容可能な訂正が施せた．とくに，大規模文書から構築した言語モデルの確率と，誤り文ペアの二値素性を併用することにより，再現率を大幅に向上させることができた．

参考文献

Kenji Imamura, Tomoko Izumi, Kugatsu Sadamitsu, Kuniko Saito, Satoshi Kobashikawa, and Hirokazu Masataki. 2011. Morpheme conversion for connecting speech recognizer and language analyzers in unsegmented languages. In *Proc. of Interspeech 2011*, pages 1405–1408.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the ICML-2001*, pages 282–289.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated japanese error correction of second language learners. In *Proc. of IJCNLP 2011*, pages 147–155.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proc. of ASRU 2011*.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proc. of ACL-08: HLT*, pages 665–673.

Hisami Suzuki and Kristina Toutanova. 2006. Learning to predict case markers in japanese. In *Proc. of ACL-COLING 2006*, pages 1049–1056.

大木 環美, 大山 浩美, 北内 啓, 末永 高志, 松本 裕治. 2011. 非日本語母国語話者の作成するシステム開発文書を対象とした助詞の誤用判定. **言語処理学会第 17 回年次大会発表論文集**, pages 1047–1050.