3I1-R-9-7

文字–単語アライメントを用いた日本語学習者の作文誤り訂正

Error Correction for Japanese as a Second Language Learners' Text Using Character-to-Word Alignment

水本 智也*1 Tomoya Mizumoto　小町 守*1 Mamoru Komachi　永田 昌明*2 Masaaki Nagata　松本 裕治*1 Yuji Matsumoto

*1奈良先端科学技術大学院大学 Nara Institute of Science and Technology　*2NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

Recently, natural language processing research has begun to pay attention to second language learning. Statistical machine translation-based approach has been proposed to correct all types of errors. However, word-to-word alignment does not perform well on Japanese learners' texts since they are hard to tokenize. To solve this problem, we proposed character-to-character alignment model but it fails to use word information explicitly. In this paper, we extend the character-to-character alignment model to character-to-word alignment model. It uses word information on the target side, so it is able to use more linguistic information such as word n-gram language model than character-to-character alignment model. We show that the character-to-word alignment model achieves higher f-measure than the word-to-word and the character-to-character alignment model.

1. はじめに

近年，自然言語処理による第 2 言語学習支援に関する研究が注目を集めている [Rozovskaya 11, Liu 11, Mizumoto 11, Oyama 10]．たとえば，日本語学習者の作文誤り訂正に関する研究では，[Oyama 10, 南保 07, 今枝 03] が助詞の誤り訂正を行なっている．図 1 に NAIST 誤用コーパス [大山 09] から調べた，日本語学習者の誤り傾向を示す．助詞が 24%を占めており，日本語学習者の誤りやすい箇所だとわかるが，残りの 76%は語彙選択や表記など助詞以外の誤りであるため，これらの誤りの検出や訂正も重要である．

[Brockett 06] は統計的機械翻訳の手法を用いて，英語の加算・不加算名詞を対象とした誤り訂正を行なった．しかしながら，日本語文は単語境界が明示的でないため，誤りを含んでいたり，ひらがなで書かれていたりすると単語分割に失敗してしまう問題がある．[Brockett 06] は英語を対象としていたため，単語分割を失敗するという問題はなかった．そこで [Mizumoto 11] は，日本語を対象とした時に問題となる単語分割の問題を解決するため，学習者コーパスの学習者の書いた文と添削後の文の両方を文字単位に分割して，文字単位の対応をとる（文字–文字アライメント）手法を提案した．

しかしながら，[Mizumoto 11] の文字–文字アライメントによる誤り訂正手法では，辞書や品詞といった単語に関する情報を十分に活用できないという問題点があった．そこで，本研究ではその問題を解決するために，学習者の書いた文は文字分割し，添削後の文は単語分割を行なうことで，文字列と単語列の対応をとる（文字–単語アライメント）手法を提案する．また，各手法で訂正できる誤りの種類が異なるという傾向を利用して，各手法を組み合わせて訂正を行なうことでさらなる性能改善ができることを示す．

以下，2 節でフレーズベースの統計的機械翻訳を使った日本語誤り訂正手法について説明を行ない，3 節で提案手法である文字–単語アライメントを用いた誤り訂正の実験とその結果および考察を行なう．その考察を元に得た，各手法で訂正できる誤りの種類が異なるという傾向を利用して，各手法を組み合わせて訂正を行なった実験と考察に関して 4 節で述べる．

連絡先: 水本智也，奈良先端科学技術大学院大学　情報科学研究科，奈良県生駒市高山町 8916-5 情報科学研究科　自然言語処理学研究室，tomoya-m@is.naist.jp

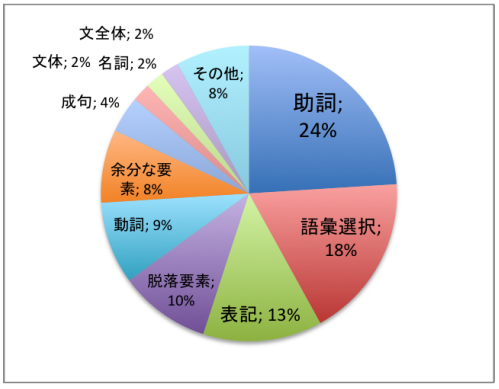


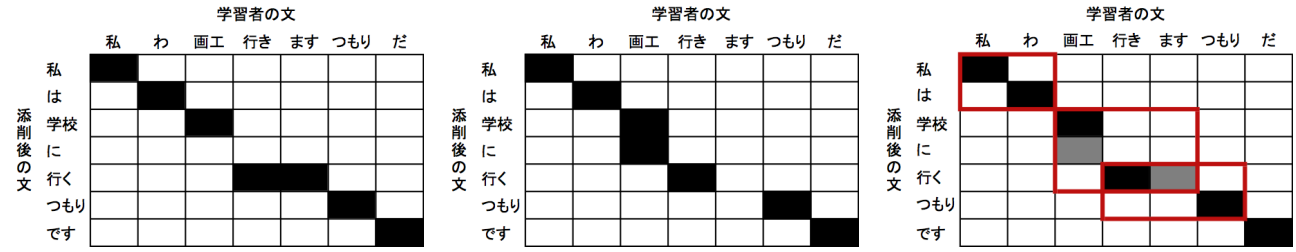
図 1: 日本語学習者の作文誤り傾向

2. フレーズベース統計的機械翻訳を使った日本語学習者の作文誤り訂正

本研究ではフレーズベースの統計的機械翻訳を用いて日本語学習者の作文誤り訂正を行う．

$$\hat{e} = \arg\max_e P(e|f) = \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(e,f) \qquad (1)$$

式 1 は対数線形モデルを使った統計的機械翻訳の式である．ここで e はターゲット側（翻訳後の言語）であり，f がソース側（翻訳前の言語）である．$h_m(e,f)$ は M 個の素性関数であり，λ_m が各素性関数に対する重みである．ソース側の文 f に対して，素性関数の重み付き線形和を最大化するターゲット側の文 e を探せばいいことを意味している．素性関数には，翻訳モデルや言語モデルなどが用いられる．翻訳モデルは一般に $P(f|e)$ という条件付き確率の形で表される．この翻訳モデル $P(f|e)$ はフレーズ間の対訳確率に分解して定義される．言語モデルは一般に $P(e)$ という確率の形で表され，n-gram 言語モデルが広



(a) 学習者の文から添削後の文への単語アライメント (b) 添削後の文から学習者の文への単語アライメント (c) フレーズ抽出

図 2: 単語アライメントからのフレーズ抽出

く用いられている．また，翻訳モデルは文単位で 1 対 1 対応のとれた対訳コーパスから学習し，言語モデルはターゲット側言語から学習することができる．

これを誤り訂正に適応した場合，f は学習者の書いた添削前の文となり，e は添削された後の文となる．また，翻訳モデルは添削前後の文で 1 対 1 対応のとれた添削コーパスから学習し，言語モデルは正しい文から学習したものになる．こうすることで統計的機械翻訳を使って，誤りを含む文を正しい文に変換することができる．

2.1 フレーズ抽出方法

フレーズの抽出は任意のアライメントのとれた対訳コーパスからヒューリスティックスを用いて行なう．ここでは単語アライメントの例を用いて説明を行なう．フレーズ抽出を行なう前処理として，はじめに “学習者の書いた文から添削後の文” と “添削後の文から学習者の書いた文” の両方向の単語のアライメントをとる．学習者の書いた文から添削後の文への単語アライメントの結果を図 2(a) に，添削後の文から学習者の文への単語アライメントの結果を図 2(b) に示す．単語のアライメントには，統計的機械翻訳で広く使われている GIZA++（HMM と IBM モデルを使用）[*1]を使用する．GIZA++のヒューリスティックスでは，両者の積集合を求めその積集合の対応点を起点とし，和集合の中の対応点から既にある対応点に隣接する点を加えていく．図 2(c) の黒の部分が積集合の点で，灰色の部分が新たに加えられた点である．その後，対応点が内部に閉じているようなフレーズの対応を取り出す．赤の線で囲んである部分が取り出されるフレーズの一部であり，“私わ” を “私は” に，“画工行きます” を “学校に行く” に，“行きますつもり” を “行くつもり” に対応させるフレーズを抽出することができる．

2.2 文字–文字アライメントを用いたフレーズ抽出

まず，[Mizumoto 11] が行なった文字–文字アライメントを用いた手法について説明を行なう．一般的に機械翻訳で日本語から他の言語へ変換する場合，単語に分割してから行うが，学習者の書いた文は新聞記事などの日本語母語話者の書いた編集済みのコーパスから学習された形態素解析器では，単語にうまく分割することができない．たとえば，

でもじょずじゃりません

といった学習者の書いた文があるが，これを MeCab [*2] を用いて分かち書きを行うと，

でも じ ょずじゃりません

と「ょずじゃりません」が未知語として分割されてしまう．一方，実際に人が添削した後の文を MeCab で分かち書きすると，

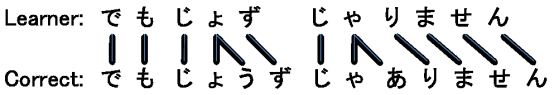


図 3: 文字–文字アライメントの例

でも じょうず じゃ あり ませ ん

のようになる．この 2 つを見ると，学習者の文には解析誤りが含まれており，添削後の文と対応を取ることができないため，図 2 のように適切なフレーズをとることが難しく，単純に統計的機械翻訳を用いてフレーズ抽出することは困難であると予想される．

そこで，[Mizumoto 11] は単語よりも細かい文字の単位に分割することを提案した．文字分割にすることにより形態素解析器の誤りの影響を受けないため，頑健な解析が可能である．実際に上記の学習者の例と正解の例を文字に分割し，アライメントをとった結果が図 3 のようになる．フレーズベースの手法を用いることで,「じょず」と「じょうず」との対応を学習でき，単語分割した場合と比べると頑健な誤り訂正が期待できる．

2.3 文字–単語アライメントを用いたフレーズ抽出

一方，[Mizumoto 11] では学習者の文も添削後の文の両方とも文字分割を行なっているため，単語の情報が失われてしまっている．学習者の作文は単語分割に失敗することがあるが，添削後の文は正しく単語分割できる可能性が高いにもかかわらず，文字-文字アライメントを用いた手法では単語の情報を十分に活用することができない[*3]．

そこで，学習者の文は文字分割，正しい文は単語分割にした “文字–単語アライメント” を用いたフレーズ抽出手法を提案する．図 4 は文字–単語アライメントの例である．“じょず” から “じょうず” への変換を学習できるのはもちろん，“じょずじゃ” から “じょうずじゃ” への変換を学習することができる．学習者の文は文字に分割することで未知語の問題を軽減し，添削後の文は単語に分割することで品詞などのリッチな情報を活用することができる．これにより学習者の文も正しい文に対しても文字分割を行なった場合よりも，正しく訂正できると考えられる．

3. 誤り訂正の分割単位による比較実験

提案した文字–単語アライメントモデルを用いた統計的機械翻訳による日本語学習者の作文誤り訂正性能を評価するために実験を行なった．実験には Moses 2010-08-13 [*4]をデコーダ，

*1 http://code.google.com/p/giza-pp/
*2 http://mecab.sourceforge.net/
*3 学習者コーパスに辞書のエントリを加えるなど，辞書の情報を間接的に入れることも可能ではある．
*4 http://http://www.statmt.org/moses/

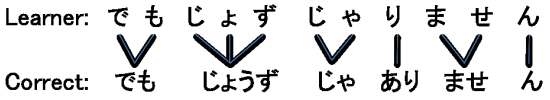


図 4: 文字–単語アライメントの例

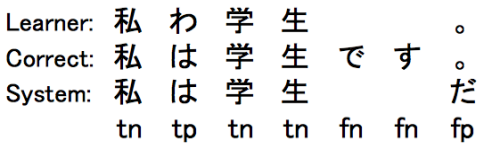


図 5: 文字単位の誤り訂正の再現率・適合率の評価

GIZA++ 1.0.5 [*5]をアライメントのツールとして利用した．また，単語分割には UniDic 1.3.12 [*6]を辞書として用いた MeCab 0.97 を利用した．

ベースラインとして単語–単語アライメントモデル，文字–文字アライメントモデルの 2 種類を準備した．単語–単語アライメントモデルおよび文字–単語アライメントモデルの言語モデルには単語 3-gram を使用し，文字–文字アライメントモデルの言語モデルには文字 5-gram を使用した．全てのモデルに対して，minimum error rate training (MERT) [Och 03] を行なった．

3.1 実験データ

実験には，クローリングを行ない獲得した 2010 年 12 月までの言語学習 SNS の Lang-8 [*7]のデータを用いた．Lang-8 は学習者同士で相互に作文を添削しあうサイトであり，学習者の書いた文とその添削文が対になった大規模なデータを手に入れることができる．

トレーニングデータとして日本語学習者の書いた作文とその添削後の文 796,956 文対を用いた．テストおよびデベロップメントデータには，トレーニングデータとは別に，英語を母語とする学習者が書いた作文 500 文を用意して人手で再添削を行なったものを使用した．500 文のうち 200 文をテスト，残りの 300 文をデベロップメントデータとして利用した．

3.2 評価尺度

評価尺度として，文字単位による再現率 (R)，適合率 (P) および F 値 (F) を用いた．再現率，適合率，F 値は以下のように定義する．

$$\text{再現率} = \frac{tp}{tp+fn}, \qquad \text{適合率} = \frac{tp}{tp+fp},$$
$$F\ \text{値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

tp (true positive) はシステムが訂正を行ない正解だった箇所，fp (false positive) はシステムが訂正を行なったが訂正する必要がなかった箇所，fn (false negative) はシステムは訂正を行なわなかったが訂正が必要だった箇所である．また，F 値は再現率と適合率の調和平均である．図 5 の例を用いて説明を行なう．図 5 中の tn はシステムが訂正を行なわず正解だった箇所である．tp の数は 1，fp の数は 1，fn の数は 2 である．したがって，再現率は 1/3 となり，適合率は 1/2 となる．

*5 http://code.google.com/p/giza-pp/
*6 http://www.tokuteicorpus.jp/dist/
*7 http://lang-8.com/

表 1: 分割単位による比較実験結果

	再現率	適合率	F 値
単語–単語	0.116	0.348	0.172
文字–文字	0.106	**0.397**	0.166
文字–単語	**0.127**	0.313	**0.180**

3.3 実験結果

実験は 796,956 文から 790,000 文を 10 回ランダムに抽出して翻訳モデルの学習を行ない，その平均値で評価を行なった．言語モデルには，796,956 文全てを用いた．表 1 に実験の結果を示す．提案手法である文字–単語アライメントモデルが F 値，再現率がもっとも高かった．適合率に関しては，文字–単語アライメントモデルがもっとも低い値となっており，文字–文字アライメントモデルがもっとも高い値となった．

3.4 アライメントモデルによる結果の違いに関する考察

この節では，各モデルによる訂正結果の違いの例を示しながら考察を行なう．表 2 に実際の出力例を示す．表 2(a) にどのモデルを用いた場合でも訂正できた例を示す．最初の例では，“は” を “わ” に間違えるという誤りを学習者が間違えやすい誤りの典型であるため，訂正できたと考えられる．また，“おいしいだね” も学習者が間違いやすい誤りであり，どのモデルのフレーズテーブルでも “おいしいだね” から “おいしいね” への変換の対応が存在していた．

表 2(b) は単語–単語アライメントモデルの出力結果の例である．表に示しているような例で，学習者の文で誤りを含む場合でも単語分割ができる場合は，単語の対応をとることができるため訂正できた．人手による正解データと比べると “協会” を “教会” に変換しているが，これは “協会” を “教会” に変換するものがフレーズテーブルに存在しており，“協会に行く” は言語モデルにないが “教会に行く” は存在したためであると考える．“勉強” と “続けたくて” の間に “を” を挿入できなかった．これは，“勉強続き” を “勉強を続け” に変換するようなものがフレーズテーブルになかったためである．

文字–文字アライメントモデルの出力例を 2(c) に示す．文字–文字アライメントモデルでは，“アバイン” を “アーバイン”，“インタネット” を “インターネット” や “けだ” を “だけ” に訂正するなど，単語のスペル誤りを訂正できた．このように文字–文字アライメントモデルはスペル誤りの訂正に強い．その理由としては，文字という細かい単位を利用をすることで，フレーズテーブルに単語として存在しない場合でも訂正が可能であるからだと考えられる．例えば，“アバイン” を “アーバイン” に変換するようなものはどのモデルでもフレーズテーブルには存在していなかった．しかしながら，文字–文字アライメントモデルのフレーズテーブルを見ると，“のア” を “のアー” の対応は存在しており，訂正する際にこれと言語モデルを用いることで訂正できたと推測される．

文字–単語アライメントモデルのみで訂正できた例を表 2(d) に示す．完全に訂正はできていないが “祖ごく” を “祖すごく” というように訂正できた．これはフレーズテーブルで “祖ごく” を “すごく” に変換するものがないので，“祖” が残ってしまったと考えられる．また，表 2(d) には書いていないが，文字–単語アライメントモデルで訂正できたものを見ると，単語–単語アライメントモデルと同じものが多く，文字–文字アライメントモデルで訂正できたものも一部訂正できた．表 2(b) に示した 2 つの例や，2(c) で示した “けだ” を “だけ” に直すものは文字–単語アライメントモデルで訂正できた．

表 2: 各モデルの出力例

(a) 全てのモデルで訂正できた例

学習者の作文	TRUTHわ 美しいです
人手による添削	TRUTHは 美しいです

学習者の作文	おいしい だ ね
人手による添削	おいしい ね

(b) 単語–単語アライメントモデルの出力例

学習者の書いた文	私は協会に行 きます つもりです
システムの出力	私は教会に行 く つもりです
人手による添削	私は協会に行 く つもりです

学習者の書いた文	勉強続 き たくて
システムの出力	勉強続けたくて
人手による添削	勉強 を 続けたくて

(c) 文字–文字アライメントモデル出力例

学習者の書いた文	カルフォルニアのアバイン大学
システムの出力	カルフォルニアのア ー バイン大学
人手による添削	カルフォルニア大学のア ー バイン校

学習者の書いた文	インタネットを壊した
システムの出力	インタ ー ネットを壊した
人手による添削	インタ ー ネットを壊した

学習者の書いた文	いつも英語 けだ を話したくない
システムの出力	いつも英語 だけ を話したくない
人手による添削	いつも英語 だけ を話したくない

(d) 文字–単語アライメントモデルの出力例

学習者の書いた文	祖 ごく忙しかった
システムの出力	祖す ごく忙しかった
人手による添削	す ごく忙しかった

4. 繰り返しによる誤り訂正実験および考察

3.4 節で各モデルの結果の違いの考察を行なったが，そこで各モデルでそれぞれ訂正できる誤りが異なっているということがわかった．このことから，一度あるモデルで訂正を行なった後に別のモデルでもう一度訂正を行なうことで，訂正できていなかった誤りを新たに訂正できるのではないかという仮説を立てることができる．

そこで，この仮説を確かめるために比較実験を行なった．行なう実験は文字–文字アライメントモデルと文字–単語アライメントモデルのどちらを 1 回目と 2 回目に使うか，2 × 2 の合計 4 つである．

実験の結果を表 3 に示す．“文字–文字 → 文字–文字” も “文字–単語 → 文字–単語” のような同じアライメントモデルを繰り返すものでも 1 回訂正した場合よりも F 値は高くなった．これは，一度訂正を行なったことにより二度目の訂正の時にフレーズテーブルで一致する箇所が出てくるためだと考えられる．表 2(b) の例を使って具体的に説明を行なうと，“勉強続き” を “勉強を続け” に変換するようなものがフレーズテーブルになかったとしても，“続きたくて” を “続けたくて” に，“勉強続け” を “勉強を続け” に変換するようなものがフレーズテーブルにあれば，一度目で “続けたくて” に訂正し，二度目で “勉強を続け” と訂正が可能になるためである．

表 3: 2 度の繰り返しを行なった実験結果

	再現率	適合率	F 値
文字–文字 → 文字–文字	0.119	**0.367**	0.179
文字–単語 → 文字–単語	0.143	0.295	0.192
文字–単語 → 文字–文字	0.160	0.319	0.212
文字–文字 → 文字–単語	**0.165**	0.356	**0.225**

同じモデルを 2 回繰り返すよりも，“文字–単語 → 文字–文字” や “文字–文字 → 文字–単語” のように違うアライメントモデルを用いた場合の方が F 値が高くなっている．これは仮説で述べたように一度目で訂正した誤りとは異なる種類の誤りを訂正できるためであると考える．実際，3.4 の例で挙げた “教会に行きますつもりです” は文字–文字アライメントモデル，“文字–文字 → 文字–文字” では訂正できないが “文字–文字 → 文字–単語” では訂正できた．

5. おわりに

本研究では，“文字–単語アライメント” を用いた日本語学習者の作文誤り訂正の手法を提案した．実験を行なった結果，文字–単語アライメントモデルの F 値がもっとも高く，誤りの種類によって有効なモデルが異なるということがわかった．訂正できる誤りが異なっているということを利用して，あるモデルで 1 度訂正を行なった後，異なるモデルでもう 1 度訂正を行なうという実験を行なった．その結果，異なるモデルを組み合わせて用いることでさらなる性能改善ができることがわかった．

参考文献

[Brockett 06] Brockett, C., Dolan, W. B., and Gamon, M.: Correcting ESL Errors Using Phrasal SMT Techniques, in *Proceedings of COLING-ACL*, pp. 249–256 (2006)

[今枝 03] 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 桝井文人: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研究会報告, pp. 39–46 (2003)

[Liu 11] Liu, X., Han, B., and Zhou, M.: Correcting Verb Selection Errors for ESL with the Perceptron, in *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pp. 411–423, Berlin, Heidelberg (2011), Springer-Verlag

[Mizumoto 11] Mizumoto, T., Komachi, M., Nagata, M., and Matsumoto, Y.: Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners, in *Proceedings of IJCNLP*, pp. 147–155 (2011)

[南保 07] 南保亮太, 乙武北斗, 荒木健治:文節内の特徴を用いた日本語助詞誤りの自動検出・校正, 情報処理学会研究報告 自然言語処理研究報告, pp. 107–112, 情報処理学会 (2007)

[Och 03] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, in *Proceedings of ACL*, pp. 160–167 (2003)

[大山 09] 大山浩美:日本語学習者コーパスのための誤用タグ構築について, 熊本県立大学日本語日本文学会, 54, pp. 102–114 (2009)

[Oyama 10] Oyama, H. and Matsumoto, Y.: Automatic Error Detection Method for Japanese Case Particles in Japanese Language Learners, in *Corpus, ICT, and Language Education*, pp. 235–245 (2010)

[Rozovskaya 11] Rozovskaya, A. and Roth, D.: Algorithm Selection and Model Adaptation for ESL Correction Tasks, in *Proceedings of ACL*, pp. 924–933 (2011)