Vol.2015-NL-223 No.2
2015/9/27

部分的アノテーションを利用したCRFによる日本語学習者文の単語分割

塘　優旗[†1,a)]　小町 守[†1,b)]

概要：日本語学習・教育支援として，学習者の書いた作文の誤り検出，誤り訂正の必要性が高まってきている．そのような技術の精度向上には，学習者の文に頑健な単語分割が重要である．本稿では，言語学習者サイトである Lang-8 における日本語学習者の日本語文とそれに対しての添削文のペアから，学習者の単語分割に関しての訓練用データを作成する．しかし，Lang-8 では添削されるべき部分の一部のみしか訂正されていない場合もあり，信頼できないデータも多い．そこで，おそらく信頼できるであろう，添削が行われた部分の文字のみに単語境界のアノテーションをすることで学習者コーパスを作成し，アノテーションが曖昧な部分に関しては周辺尤度を用いて訓練を行う条件付き確率場の拡張 [1] を利用することで訓練を行う．訓練時に利用する学習者コーパス中の文を，学習者文と添削文間での挿入，削除数によって制限し，分野適応することで，学習者テキストの単語分割精度を向上させることができることを示す．また，その他比較手法との違いを実際の出力結果を交えて考察する．

1. はじめに

国際交流基金の「2012 年度 日本語教育機関調査」によると，海外の 136 の国と地域で，399 万人の人々が日本語を学習しており，日本語学習者の数は年々増加している．一方，日本語教師の数は 6.3 万人に止まり，全ての日本語学習者が十分な学習環境を得られているわけではない．日本語教師の数は 2009 年の調査結果から 28 %増加しているが，日本語教師の不足を補うために，作文誤り検出・訂正などの自動添削を用いて日本語教師・学習者の支援をする必要性がある．

自然言語処理を利用して作文誤り検出・訂正を行うには，まず，単語分割を行う必要がある．たとえば，Mizumoto ら [2] の統計的機械翻訳の手法を用いた日本語学習者の作文自動誤り訂正においては，正しく分割できた場合は訂正の精度が高くなることが述べられている．しかしながら，日本語学習者の文は，うまく文字の変換がされていなかったり，誤りを含むなどの理由から，既存の単語分割器や形態素解析器では単語分割に失敗しやすい．また，今村ら [3] は，助詞誤りに限定し誤り訂正を行っているが，訓練の際に助詞誤りのみ含まれる学習者文と修正文のペアを用いているため，形態素解析で助詞をうまく分割できなかった場合は助詞の誤り訂正を行うことができない．したがって，日本語学習者の文をうまく単語分割できるようになることで，誤り訂正の精度向上に貢献できることが考えられる．

そこで，本研究では日本語学習者の日本語文を，誤り訂正に適した形に単語分割することを目標とする．藤野ら [4] の先行研究を参考に，日本語学習者の書いた文と添削文のペアから添削が行われた箇所のみ単語分割のアノテーションを行い，その他の箇所に関しては曖昧なままにした学習者コーパスを作成する．藤野らは学習者コーパスを用いて KyTea を追加学習することで誤り箇所の分割精度は向上することができたが，誤っていない箇所も含めた全体の精度は低下することを報告している．一方今回はアノテーションに曖昧な箇所が含まれていても系列全体を用いて訓練が行える条件付き確率場の拡張 [1] を利用することで訓練を行い，訓練用データの利用の仕方で大きく結果が異なり，全体の精度が向上することを示す．また，それに際してテストデータとして日本語学習者文の単語分割に関してのデータセットの作成を行い一致率を確認した．

2. 関連研究

現在，日本語単語分割の手法として主に利用されているのは，ルールベースのもの [*1] や，機械学習に基づくもの [5] である．これらの単語分割の手法は，誤りを含んだ文や整っていない文に対して，単語分割の精度が落ちてし

†1 首都大学東京 システムデザイン研究科 情報通信システム学域
a) tomo-yuki@ed.tmu.ac.jp
b) komachi@tmu.ac.jp

[*1] 日本語形態素解析システム JUMAN http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

まう傾向がある．これは，誤りに対応したたくさんのルールを人の手で定めることや，誤りを含む文に対して単語分割のアノテーションが行われた大量のコーパスを作ることが困難なためである．

2.1 単語分割，形態素解析のドメイン適応

誤りを含む文と同様に，ドメイン適応のためのコーパスのアノテーションはとても高コストである．そこで，ドメイン特有の箇所のみアノテーションを行う**部分的アノテーション**がなされたコーパスの利用が考えられる．Neubigら [6] は，日本語文の単語分割，読み推定タスクのドメイン適応のために部分的アノテーションにより作成されたコーパスを利用した点推定の手法を提案し，それらの実装であるKyTea [*2] を公開している．入力された文における，各単語境界，各単語に対しての読みは全て独立して推定されるため，一部の箇所のみアノテーションされたコーパスを訓練データとして利用することができる．彼らは，完全にアノテーションされた一般的なコーパスに加えて部分的アノテーションのされたコーパスを利用することで性能向上を報告している．部分的アノテーションを利用したコーパスの利用については本研究と同様であるが，コーパスのアノテーションを人手で行っている点と学習に用いる手法が異なる．

また，日本語学習者の文同様に，表記ゆれの大きいWEB上の個人が書いたマイクロブログ等のテキストに対して頑健な形態素解析を行う研究も多くなされている．笹野ら [7] は，辞書中に存在するような正規語から派生する未知表記，未知オノマトペを扱うために，単語の派生ルールとコストを人手で定め，形態素解析の際のラティス展開を拡張し性能向上を図っている．工藤ら [8] は，Web上に頻出するひらがな交じり文に対して頑健な形態素解析を提案している．正規の文から各単語がひらがな化し，ひらがな混じり文が生成されるような過程をモデル化した生成モデルを導入し，大規模なWEBコーパス及びEMアルゴリズムによってモデルのパラメータ推定を行い，ひらがな混じり文に対して頑健な形態素解析と正規文の導出を可能にしている．斉藤ら [9] は，工藤らの手法をひらがな化以外にも拡張し，単語の派生ルールを用いて正規語から辞書中に存在しない未知語に派生してしまう確率である崩れ表記語生成確率を大量の平文を用いてあらかじめ学習する．また，その崩れ表記語生成確率を素性として導入し，正解付きコーパスで学習を行っており性能向上が報告されている．これらの研究は，あらかじめ正規語からの派生パターンを与えているのに対して，本研究では直接コーパスから誤りの含まれる箇所に関しての単語分割を学習する点が異なる．

[*2] http://www.phontron.com/kytea/index-ja.html

2.2 日本語学習者の作文の誤り訂正に向けた単語分割

藤野ら [4] は言語学習者向けの相互添削型SNSであるLang-8[*3] から日本語学習者の文とそれに対応した添削文を取得し，文対間で単語分割のアノテーションを自動で行い学習者コーパスを作成することを提案している．彼らは，添削前後で変化のあった箇所の単語分割を利用するパターンを点推定の単語分割器であるKyTea-0.3.2 [6] に再学習させることが学習者の文中の誤りを含む箇所における単語分割精度を高める一方，正解箇所の精度を低くすることを報告している．本研究では，藤野ら同様に学習者コーパスを作成し，添削前後で変化のあった箇所の単語分割のみアノテーションされた文を利用して，添削前後で削除，挿入が行われた学習者コーパス中の文の影響を確認し，条件付き確率場によって訓練することで，全体の評価値が向上するような学習者コーパスの利用制限を行うことができた．

学習者の文は，誤りを含んでいるので単語分割が失敗する傾向があるが，添削文に関しては学習者の文に対して誤り訂正などの添削が行われた文であるので，一般的な単語分割の手法でうまく単語分割が可能だと仮定する．このことから，添削文の単語分割を行い，その単語境界を学習者の文に反映することで大量の学習者の単語分割コーパスを作成する．

しかし，Lang-8中の添削文の多くは，添削対象の文の言語が母語であるが必ずしも専門的知識を持たない一般のユーザの投稿である．そのため，学習者の文に複数の誤りが含まれていた場合に添削文中で訂正の不足が起こっていたり，曖昧な場合にはそのままになっていたりすることが起きうる．従って，比較的信頼することのできる誤り訂正が行われた箇所のみアノテーションを採用し，これ以外の箇所に関してはアノテーションを行わずに曖昧な状態のままとする．

このように，アノテーションを行う難しい箇所や曖昧な箇所にはアノテーションを行わずに，一部の箇所のみにアノテーションを行うことで2.1項で挙げた部分的アノテーションを実現する．以下のようなステップで行う．

(1) 添削前後の挿入，削除数の導出
学習者文と添削文のペア間において，文字の挿入・削除操作の箇所を動的計画法を用いて求める．各文字に対して，挿入 (Insertion) をIタグ，削除 (Deletion) をDタグ，操作なしをNで表すと以下のようになる．
学習者の文 でもじよ ずじゃ りません
添削文 でもじ ょうずじゃありません
文字操作タグ N N N I D D N N N D N N N N

(2) 添削文の単語分割
添削文に対して単語分割器で分割を行う．単語の開始文字をBタグ，単語の内部文字をIタグ，1文字単語

[*3] http://lang-8.com/

をSタグで表すと以下のようになる．

学習者の文　でもじよ　ずじゃ　りません

添削文　でもじ　ょうずじゃありません

単語分割（添）B I B　I I I B I B I B I S

文字操作タグ　N N N I D D N N N D N N N N

(3) 学習者文への単語分割の反映

添削文の分割箇所を学習者の文に反映する．挿入された箇所は，添削文において挿入文字の前と同じ単語を構成し，添削文で単語分割の終了文字となっていた場合，学習者文は挿入箇所の前で単語分割を行う．また，挿入文字の後ろと同じ単語を構成していた場合，学習者文はその単語の分割に従う．削除箇所は，削除箇所の次の文字から別単語になっていた場合は，削除文字を削除箇所の前に接続する．また，削除箇所の前後の文字が同じ単語を構成していた場合はその単語の分割箇所で分割を行う．したがって以下のようなアノテーションとなる．

学習者の文　でもじよずじゃりません

単語分割（学）B I B I I B I S B I S

また部分的アノテーションとした場合，添削が行われた箇所のみのアノテーションを行い曖昧なものを？タグとすると以下のようなタグ付けとなる．

学習者の文　でもじよずじゃりません

単語分割（学）？ ？ B I I ？ ？ S ？ ？ ？

2.3 部分的アノテーションを利用した条件付き確率場の学習

部分的アノテーションを行い，曖昧なアノテーションを含む学習コーパスを利用した機械学習の手法として，坪井ら [1] は，アノテーションが曖昧な部分に関しては周辺尤度を用いて学習を行う条件付き確率場（CRF）[10] の拡張を提案している．

入力列 $x = (x_1, x_2, ..., x_T)$ を入力変数 $x_t \in X$ が要素となる列構造，ラベル列 $y = (y_1, y_2, ..., y_T)$ をラベル変数 $y_t \in Y$ の列，$\Phi(x, y) : X \times Y \to \mathbf{R}^d$ を入力列 x とラベル列 y の組から d 次元の任意の素性ベクトルへの写像，$\theta \in \mathbf{R}^d$ をモデルのパラメータベクトルとすると，一般的な CRF は x が与えられた時の y の条件付確率を式 1 でモデル化する．分母は正規化項である．

$$P_\theta(y|x) = \frac{e^{\theta \cdot \Phi(x,y)}}{Z_{\theta,x,y}} \quad (1)$$

例えば，文中の単語の品詞推定タスクを系列ラベリング問題として解く場合，入力列 x は 1 文の単語列，ラベル列 y が各単語に対しての品詞列となる．

次に y の一部が曖昧なデータの表現のために，$L = (L_1, L_2, ..., L_T)$ を入力 x の各文字文字が取り得るラベル変数の値集合 $L_t \in 2^Y - \{\emptyset\}$ の列とする．例えば，PennTreebank コーパス [*4] における品詞に曖昧なアノテーションがされた文は以下のようなものがあり，“pending” の品詞は “VBG” または “JJ” としてアノテーションされている．このとき各ラベルはそれぞれ，DT：限定詞，NN：名詞単数，VBZ：動詞 3 人称単数現在形，VBG：動名詞または動詞現在分詞，JJ：形容詞，SYM：記号 である．

入力文　That　suit　is　pending .

$$L = (\{DT\}, \{NN\}, \{VBZ\}, \{VBG, JJ\}, \{SYM\}) \quad (2)$$

になる．一般的な CRF では，このように y の一部だけが曖昧な L から直接学習することができないため，L に適合するあらゆるラベル列の集合を Y_L としたとき，以下のようなモデルを用いる．

$$P_\theta(Y_L|x) = \sum_{y \in Y_L} P_\theta(y|x) \quad (3)$$

坪井らは部分的アノテーションを利用することで，ドメイン固有の表現に対応したコーパスを低コストで作成することができ，完全なアノテーションをする場合に比べて単語分割性能を向上させることを示している．

また，Liu ら [11] は日本語と同様に単語境界のない言語である中国語の単語分割に部分的アノテーションを使用した CRF を適用し性能向上を示した．また，彼らは crfsuite [12] を部分的アノテーションを利用できるように改良した partial-crfsuite [*5] を公開している．

3. 部分的アノテーションを用いた日本語学習者文の単語分割

本研究では，日本語学習者の書いた日本語文に最適化された単語分割を提案する．Lang-8 から抽出した学習者の文と添削文のペアから部分的アノテーションによって学習者コーパスを自動で作成し，完全にアノテーションされた一般的なコーパスと合わせて，坪井らの提案する CRF の訓練用データとして用いる．

3.1 問題設定

今回の日本語学習者の文の単語分割は，対象文の各文字に対して，単語開始文字 (B)，単語内文字 (I)，1 文字単語 (S) のいずれかをラベル付けする系列ラベリング問題として扱い，単語分割を行う．

3.2 単語分割基準

MeCab 等で利用される辞書である UniDic で採用されて

[*4] https://www.cis.upenn.edu/~treebank/
[*5] https://github.com/ExpResults/partial-crfsuite

いる短単位を単語分割の基準とする．短単位は国立国語研究所が規定したものであるが，UniDic で採用されているものは多少異なる．国立国語研究所の短単位では、意志・推量の助動詞「う」「よう」を独立した語として扱うが，UniDic における短単位では，これらを活用語尾とみなす．従って「でしょ/助動詞 う/助動詞」，「食べよ/動詞 う/助動詞」のような単語の連接を UniDic の基準では「でしょう/助動詞」，「食べよう/動詞」のように 1 単語としてみなし「意志推量形」という活用として扱う．

現代日本書き言葉均衡コーパス（BCCWJ）[*6] において採用されている短単位を単語分割の基準とする．

誤りが含まれている場合は，基本的に訂正を行った正しい文を単語分割した結果にならう．例えば，単語内部に余分な文字が含まれていたり，文字の順番がおかしいもの，単語の右隣を切り出しても意味を成さない余分な文字などに関しては一つの単語として見る．また，余分な助詞などの単語として認識できるものが含まれている場合は，切り出す．

以下に誤りを含む文の単語分割，系列ラベリングの例を示す．「思うい」のように「思う」に「い」のような意味を成さない無駄な文字が含まれている場合は切り出さずに 1 単語として取り扱い単語分割を行う．

入力文　上手　じゃ　ない　と　思うい　ます
ラベル　B I　B I　B I　S　B I I　B I

また，以下のように余分だが単体で意味を成す文字「が」のような単語があった場合は，前の文字「誰」から切り出す．

入力文　H G　は　誰　が　です　か　？
ラベル　B I　S　S　S　B I　S　S

4. 実験

日本語学習者文，添削文のペアから部分的アノテーションを行って学習者コーパスを作成し，一般的なコーパスとともに学習者に特有な単語分割を部分的アノテーションを用いた CRF を用いて訓練を行う．また，訓練時に利用する学習者コーパス中の文を添削前後の挿入，削除数で制限を行い，各評価値，出力結果を確認，考察を行う．

4.1 ベースライン

提案手法との比較のためのベースラインとして，点推定を用いた単語分割器 KyTea-0.4.7 を利用する．今回 KyTea を利用する理由としては，CRF 同様に辞書を必要としない単語分割が可能であり，藤野らもベースラインとして利用しているためである．使用するモデルは，KyTea に標準で付属し BCCWJ，UniDic を主に用いて構築された**デフォルトモデル**，共に配布されている BCCWJ，UniDic 等を用いて構築された**高性能 SVM モデル** [*7]，デフォルトモデルと同様のモデルを再学習できる素性ファイル kytea-0.4.2.feat[*8] と共にベースライン部分的アノテーションの学習者コーパスによって学習した**追加学習モデル**，BCCWJ のみで訓練を行った **BCCWJ モデル**，BCCWJ と学習者コーパスによって訓練した**再学習モデル**である．KyTea の設定はデフォルトの L2 正則化された SVM を利用し，窓幅を 3，文字・文字種 n-gram の上限を 3 とし訓練を行った．学習者コーパスを利用する際には，部分的アノテーションのされた訓練データを利用できるオプションを利用した．

配布されているモデル，素性ファイルで学習できるモデルは KyTea 特有の超短単位の単語分割のため，「語尾」のタグ付けがされた単語は前の単語に接続する．具体的には以下のようになる．

超短単位　ご飯/名詞　を/助詞　食べ/動詞　る/語尾
語尾接続後　ご飯/名詞　を/助詞　食べ る/動詞

また，比較のために辞書として unidic-mecab-2.1.2[*9] を使った MeCab-0.996[*10] も利用した．

表 1　各データセットの文数

データセット		文数
現代日本語書き言葉均衡コーパス（BCCWJ）		59,431
学習者コーパス	ins1del0（挿入 1 以下挿入 0）	18,181
	ins2del0（挿入 2 以下挿入 0 以下）	25,190
	ins5del5（挿入 5 以下挿入 5 以下）	611,405
	ins5del5sub3（挿入 5 以下挿入 5 以下）	589,058

4.2 データセット

水本ら [13] によって作成された言語学習者の相互添削型 SNS「Lang-8」から抽出された学習者の文と添削文がペアになった添削コーパス [*11] を用いた．学習者文，添削文共にコメントなどがカッコ中に含まれたり，単語間にスペースが多くあったため，カッコ表現，スペースは除去を行った．そのような処理を行った日本語学習者文と添削文対 1,271,065 文を利用した．テストデータはこの内 500 文として，残りを訓練用データとした．

4.2.1 訓練用データ

訓練用データ中の日本語学習者の文，添削文のペアを利用して部分的アノテーションを行い，学習者の文に対応した単語分割の部分的アノテーション済み訓練用データとなる学習者コーパスを作成する．学習者文と添削文のペア間で編集距離（削除，挿入数）が大きいものは適切な単語分割のアノテーションとならない傾向があるため，CRF の

[*6] http://pj.ninjal.ac.jp/corpus_center/bccwj/morphology.html
[*7] http://www.phontron.com/kytea/download/model/jp-0.4.7-1.mod.gz
[*8] http://www.phontron.com/kytea/download/kytea-0.4.2.feat.gz
[*9] https://osdn.jp/projects/unidic/releases/58338
[*10] http://taku910.github.io/mecab/
[*11] http://cl.naist.jp/nldata/lang-8/

学習に使用する学習者コーパスを文対での削除，挿入数によって使用文数の制限を行った．藤野らは挿入数と削除数が 5 以下のパターンと，さらに挿入数，削除数の差分が 3 以下なパターンを学習者コーパスの利用制限としていた．今回はさらにいくつかの挿入数，削除数の制限を加えて実験を行った．

加えて，「現代日本語書き言葉均衡コーパス」(BCCWJ) のコアデータに短単位基準で単語分割された 59,431 文を，全ての箇所に単語分割のラベル付け（フルアノテーション）済みの訓練用データとして利用する．

表 1 に利用するデータセットの各文数を示す．学習者コーパス中の文対における挿入数 N 以下，削除数 M 以下で利用する文の制限を行ったデータセットを ins$N$del$M$ の様に表現する．加えて，挿入数，削除数の差が P 以下の場合のものを ins$N$del$M$sub$P$ のように表現する．また $N = M = 0$ となるような文対において添削前後で変化のないものは除去した．

4.2.2 テストデータ

3.2 項で説明した短単位基準に従い Lang-8 から抽出した日本語文 500 件に 2 人で単語分割のアノテーションを行った．片方のアノテーションを正解，もう片方のアノテーションをシステム出力と考えて一致率を評価した場合の F 値は 97.23%となった．また，単語分割のアノテーションに差異のあった文のうち，24.5%の文が誤り箇所に由来する差異であった．

4.3 素性テンプレート

CRF の学習に際して，着目する文字の前 3 文字，後ろ 2 文字を着目する窓として，文字 1，2，3-gram，文字種 1，2，3-gram を素性として用いる．具体的には以下のように文が入力された場合に，“ソ” に着目しラベル付けを行うとすると下記のように素性が選択される．

入力文　中 国 で サ [ソ] リ を 食 べ る .

文字 1，2，3-gram

1-gram : “国”, “で”, “サ”, “ソ”, “リ”, “を”

2-gram : “国で”, “でサ”, “サソ”, “ソリ”, “リを”

3-gram : “国でサ”, “でサソ”, “サソリ”, “ソリを”

文字種 1，2，3-gram

1-gram : “漢字”, “ひらがな”, “カタカナ”, …

2-gram : “漢字/ひらがな”, “ひらがな/カタカナ”, “カタカナ/カタカナ”, …

3-gram : “漢字/ひらがな/カタカナ”, “ひらがな/カタカナ/カタカナ”, …

4.4 評価手法

単語分割の評価手法として，conlleval.pl [*12] の評価スクリプトを利用して，システムから出力された単語分割結果に対して，適合率，再現率，F 値を導出する．正解文に含まれる総単語数を N_{REF}， システムの単語分割の結果に含まれる総単語数を N_{SYS}，システムの出力のうち正解文中の単語と一致するものを N_{COR} とすると，適合率は N_{COR}/N_{SYS}，再現率は N_{COR}/N_{REF} と定義される．また，F 値は適合率と再現率の調和平均であり，適合率を P，再現率を R とすると，$2 \times P \times R \times (P + R)$ と定義される．各評価値の算出の具体例を示す．

正解コーパス　でも　じよず　じゃ　り　ませ　ん

単語分割結果　でも　じよ　ず　じゃり　ませ　ん

上記の例文の場合は，正解文の単語数が $N_{REF} = 6$，システムの単語分割結果の単語数が $N_{SYS} = 6$，分割が成功している単語は「でも」，「ませ」，「ん」のため $N_{COR} = 3$ となり，適合率は $N_{COR}/N_{SYS} = 3/6 = 1/2$ となり，適合率は $N_{COR}/N_{REF} = 3/6 = 1/2$，F 値は $2 \times P \times R \times (P + R) = 2 \times (1/2) \times (1/2) \times (1/2 + 1/2) = 1/2$ となる．

4.5 ツール

添削文の単語分割結果を学習者文に反映する際に **jpair** [*13] を利用し，添削前後で変化のあった部分の単語分割のみアノテーションを自動で行う．また，訓練，結果の出力に関しては部分的アノテーションを用いた CRF の実装である **partial-crfsuite** を利用した．

4.6 実験結果

表 2 に今回の実験結果の各評価値を示す．また，以下各手法，使用する訓練用データの組み合わせを **手法（訓練用データ）** のように示す．また，手法に関してはそれぞれ，partial-crfsuite を利用した提案手法を **P-CRF**，KyTea-0.4.7 を利用したものを **KyTea**，MeCab-0.996 を利用したものを **MeCab** のように表す．

P-CRF (BCCWJ) に比べ，文対で挿入のみ行なわれた学習者コーパスを用いた P-CRF (BCCWJ+ins1del0)，P-CRF (BCCWJ+ins2del0) は各評価値が向上した．しかし，挿入数を 1 から 2 にすることで評価値が低下する．一方，文対で削除のみ行なわれた学習者コーパスを用いた P-CRF (BCCWJ+ins0del1) に関しては，評価値が低下した．

KyTea の各種モデルの評価値を確認する．P-CRF と同様に BCCWJ，ins1del0 を用いて学習したモデルである KyTea (BCCWJ+ins1del0) は，BCCWJ のみで学習した KyTea (BCCWJ) に比べ各評価値が低下した．また，

[*12] http://www.cnts.ua.ac.be/conll2000/chunking/output.html
[*13] https://github.com/tkyf/jpair

表 2 日本語学習者文の単語分割における各手法および訓練データの比較

手法	訓練用データ	適合率 (%)	再現率 (%)	F 値 (%)	フレーズ数	正解フレーズ数
P-CRF	BCCWJ のみ	95.67	97.12	96.39	8,493	8,125
	BCCWJ + ins1del0	97.31	97.65	97.48	8,395	8,169
	BCCWJ + ins2del0	96.90	97.42	97.16	8,411	8,150
	BCCWJ + ins0del1	94.51	90.59	92.51	8,019	7,579
KyTea	BCCWJ のみ	**97.44**	97.38	97.41	8,409	8,118
	BCCWJ + ins1del0	97.05	97.29	97.17	8,386	8,139
	デフォルト (BCCWJ+Unidic)	96.43	96.89	96.66	8,406	8,106
	高性能 SVM (BCCWJ+Unidic)	96.54	97.04	96.79	8,409	8,118
	素性ファイルのみ	90.34	91.26	90.80	8,451	7,635
	素性ファイル + ins1del0	90.40	91.32	90.86	8,451	7,640
	素性ファイル + ins2del0	90.42	91.37	90.89	8,454	7,644
	素性ファイル + ins5del5	80.52	68.04	73.75	7,069	5,692
	素性ファイル + ins5del5sub3	81.12	69.13	74.64	7,129	5,783
MeCab	BCCWJ+UniDic	97.09	**98.16**	**97.62**	8,458	8,139

KyTea (BCCWJ) は適合率に関しては最も高い値となった．素性ファイルと各種学習者コーパスを用いて学習を行った追加学習モデルはデフォルトモデル，高性能 SVM モデルに比べ比較的悪い結果になった．

同じ訓練用データを利用した KyTea (BCCWJ+ins1del0) と P-CRF (BCCWJ+ins1del0) を比較すると P-CRF を利用した場合の方が KyTea より高い評価値が得られた．

5. 考察

具体的に，各モデル間で単語分割が改善，悪化した具体例を踏まえて考察を行う．また，単語分割が失敗するパターンは単語境界ではない箇所で分割してしまう “過分割” と，本来の単語境界で分割されない “未分割” に分けられる．具体的には以下の例のようなパターンである．

過分割 けんどう｜と｜から｜て ｜を｜し｜ます
未分割 待っ｜て｜ほがいい ｜です｜ね｜？

5.1 挿入数，削除数の影響

P-CRF (BCCWJ) と P-CRF (BCCWJ+ins1del0) のモデル間で改善した具体例を表 3 に示す．学習者が文字を削除し，添削で挿入操作が行なわれる必要のある箇所の単語分割が改善された．また表 4 のように P-CRF (BCCWJ) は，半角英字，記号を単語分割してしまう傾向が見られたが，過分割が改善された．表 5 には単語分割結果が悪化した例を示す．「してる」を「して｜いる」の誤りとして扱い「して｜る」のように単語分割された例がいくつか見られた．実際，書き言葉では「して｜いる」が正しいとされるので，学習者文中の「てる」のようなフレーズに対して添削文中「て｜いる」のように添削が行われている例が表 6 のように学習者コーパス中に多く確認できた．一方，テストデータの正解は UniDic を参考に「し｜てる」のような単語分割で行ったため，これは不正解の事例とされた．しかしながら，学習者の単語分割として行うことを考えると，今回の提案手法でなされる単語分割の方が適切ではないだろうか．

次に，P-CRF (BCCWJ+ins1del0) から P-CRF (BCCWJ+ins2del0) に挿入数を増やした場合に表 7 に示すような箇所の単語分割が悪化した．「思い」のように単語境界に関して左側が漢字，右側がひらがなとなるような単語をより途中で分割してしまう傾向が見られた．挿入数を 1 から 2 に増やすと，学習者コーパス中の利用する文数は 1.39 倍になったのに対して，各漢字の右側の境界で単語分割のされるアノテーション数は 1.82 倍に増加し，そのような点に影響されたのではないかと考えられる．また，改善された文は 3 文のみで，改善点はほぼ確認できなかった．

文対で削除操作が行われた文を含む学習者コーパスを使った P-CRF (BCCWJ+ins0del1) は，表 8 に示すような箇所の単語分割が悪化した．「の」,「は」のような 1 文字の助詞，助動詞を前の単語に接続してしまうような単語分割の傾向が多く見られた．これは，日本語学習者が余分にそのような単語を挿入しがちで，表 9 のように，今回の学習者コーパスの作成方法では，添削前後でそのような助詞等を前の単語と接続してしまうようなラベル付けがされてしまうためである．

5.2 KyTea との比較

同じ訓練用データを利用した KyTea (BCCWJ+ins1del0) と P-CRF (BCCWJ+ins1del0) のモデル間で出力結果を比較すると表 10 に示すような箇所が改善された．特に，削除誤りが起こり，添削によって挿入が行なわれる必要のある箇所が良い結果となった．以上のようなことから，KyTea を利用する場合に比べ，部分的アノテーションを利用した CRF を利用する場合の方がより学習者コーパスの影響が良くも悪くも大きいことがわかる．

表 3 P-CRF (BCCWJ)，P-CRF (BCCWJ+ins1del0）の削除誤り箇所の単語分割改善例

BCCWJ	BCCWJ+ins1del0
ダフト｜パンク｜は｜人気｜が｜ あた ｜と｜思い｜まし｜た｜。	ダフト｜パンク｜は｜人気｜が｜ あ｜た ｜と｜思い｜まし｜た｜。
おもしろい｜ブログ｜を｜書き｜たい｜です｜が｜なに｜も｜ がんがえら ｜ない｜。	おもしろい｜ブログ｜を｜書き｜たい｜です｜が｜なに｜も｜ がんがえ｜ら ｜ない｜。

表 4 P-CRF (BCCWJ)，P-CRF (BCCWJ+ins1del0）の半角文字箇所の単語分割改善例

P-CRF (BCCWJ)	P-CRF (BCCWJ+ins1del0)
で｜も｜、｜字幕｜が｜、｜”｜ I ｜ s ｜ h ｜ e ｜ d ｜ ea ｜ d ｜?｜”｜と｜言っ｜た｜。	で｜も｜、｜字幕｜が｜、｜”｜ Ishedead ｜?｜”｜と｜言っ｜た｜。
H I T T ｜ さ ん ｜ は ｜ 日 本 ｜ に ｜ 盛 ん ｜ (｜ p ｜ o ｜ p ｜ u ｜ la ｜ r ｜) ｜です｜か｜?	H I T T｜さん｜は｜日本｜に｜盛ん｜(｜ popular ｜) ｜です｜か｜?

また，BCCWJ 単体で学習者文の単語分割にある程度有効であることが確認できた．BCCWJ は書籍や新聞中の整った文だけでなく，ブログやネット掲示板等のいわば整っていない文を含んでいるため，学習者文の単語分割に有用であったのではないかと考えられる．

KyTea を利用したモデルで一番評価値の高かった高性能 SVM モデル，P-CRF (BCCWJ+ins1del0) 間で P-CRF の良かった例を表 11 に示す．KyTea における高性能モデルでは，「でしょう」，「やろう」などの今回採用されている短単位において意志推量系と判断される活用形が「でしょ｜う」，「よろ｜う」のように分割されてしまう傾向が確認できた．今回，正解データの単語分割基準を 3.2 項で説明したものとしたため，「でしょう」，「やろう」のような単語は助動詞，動詞の意志推量形として扱われる．一方 KyTea で採用されている単語分割基準の超短単位では，「で/動詞 しょ/語尾 う/助動詞」，「や/動詞 ろ/語尾 う/助動詞」のように単語分割がされ，今回は語尾を前の単語に接続する設定で単語分割を行っているので「でしょ｜う」，「やろ｜う」のように分割される．その他の KyTea と共に配布されているモデル，素性ファイルを用いたモデルとの比較ではこれらと同様の違いが誤りとして出てしまった．このため，提案手法との比較には不適切であった．

今回，藤野らの実験と比較して KyTea (素性ファイル+ins5del5)，KyTea (素性ファイル+ins5del5sub3) の結果が大きく性能が低下してしまった．素性ファイルが超短単位で学習されるのに対して，学習者コーパスが短単位でアノテーションされたものであったため単語分割の基準にずれが生じうまく訓練できなかったことと，テストデータを新たに短単位でアノテーションしたことが理由として考えられる．

5.3 MeCab との比較

表 12, 13 にそれぞれ，MeCab と P-CRF (ins1del0) 間で単語分割が改善した例，悪化した例を示す．ひらがなを多く含み，誤りのある箇所，変換ミスのある箇所等が改善されたことが確認できた．また，複数の語で構成され 1 語となるような固有名詞や複合動詞などを分割しすぎたり，連続する 1 文字で 1 単語を構成する漢字を接続する傾向が悪化点としてあげられる．MeCab との比較では，複数の語で構成され 1 語となるような固有名詞や複合動詞などの分割がうまくいってないことが分かったが，これは今回の手法では辞書の参照をシステム中で行っていないことに起因する．

6. おわりに

部分アノテーションを利用した CRF に，BCCWJ と学習者コーパス中の添削前後で挿入数が 1 の文を学習に利用することで，BCCWJ 単体の場合に比べ評価値が向上し，誤り箇所についてもうまく単語分割ができるようになったことを確認した．しかし，学習者コーパス中の添削前後で削除が行われた文の部分的アノテーションがうまくいっていないことが確認され，今後これらのデータを有効に利用する方法を検討中である．加えて，MeCab と比較して複合語等の単語分割がうまくいかなかったため，システム中で辞書の参照を導入したい．

今回は，4.2.1 項で示したように挿入，削除数のみで制限を行い，文数での制限を行わなかった．このことにより，挿入，削除数を増加すると使用する学習者コーパス中の学習に利用する文数が増大し，学習者コーパスの影響が大きくなることで，一般的な表現の単語分割がうまくできなくなったことが考えられる．そのため，使用する学習者コーパスの最適な文数も調査を行いたい．

また，2.1 項において示した斉藤らのように，あらかじめ学習者が誤りやすいパターンについては最初から与えて学習を行うことも有効ではないかと考えられる．

謝辞

藤野拓也様，喜洋洋様，各種データの提供ありがとうございました．

表 5 P-CRF (BCCWJ)，P-CRF (BCCWJ+ins1del0) 間の単語分割悪化例

P-CRF (BCCWJ)	P-CRF (BCCWJ+ins1del0)
今｜，｜東京｜で｜一人暮らし｜し｜ てる ｜。	今｜，｜東京｜で｜一人暮らし｜し｜ て｜る ｜。
まもり｜すずけ｜てる	まもり｜すずけ｜ て｜る

表 6 ins1del0 の学習者文と添削文のペア

学習者文	添削文
オハイオ大学に行くかどうかまだ考え てる。	オハイオ大学に行くかどうかまだ考え ている。
なぜあこがれ てる のか？	なぜあこがれ ている のか？

参考文献

[1] 坪井祐太，　森信介，鹿島久嗣，小田裕樹，松本裕治：日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習，情報処理学会論文誌，Vol. 50, No. 6, pp. 1622–1635 (2009).

[2] 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., *IJCNLP*, pp. 147–155 (2011).

[3] 今村賢治，齋藤邦子，貞光九月，　西川仁：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，Vol. 19, No. 5, pp. 381–400 (2012).

[4] 藤野拓也，水本智也，　小町守，永田昌明，松本裕治：日本語学習者の作文の誤り訂正に向けた単語分割，言語処理学会第 18 回年次大会，pp. 26–29 (2012).

[5] 　森信介，中田陽介，Neubig, G.，河原達也：点予測による形態素解析，自然言語処理，Vol. 18, No. 4, pp. 367–381 (2011).

[6] Neubig, G., Nakata, Y. and Mori, S.: Pointwise prediction for robust, adaptable Japanese morphological analysis, *ACL*, pp. 529–533 (2011).

[7] 笹野遼平，黒橋禎夫，　奥村学：日本語形態素解析における未知語処理の一手法―既知語から派生した表記と未知オノマトペの処理―，自然言語処理，Vol. 21, No. 6, pp. 1183–1205 (2014).

[8] 　工藤拓，　市川宙，Talbot, D.，賀沢秀人：Web 上のひらがな交じり文に頑健な形態素解析，言語処理学会第 18 回年次大会，pp. 1272–1275 (2012).

[9] 斉藤いつみ，貞光九月，浅野久子，松尾義博：崩れ表記語の生成確率を用いた表記正規化と形態素解析，言語処理学会第 21 回年次大会，pp. 51–54 (2015).

[10] Lafferty, J., McCallum, A. and Pereira, F. C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *ICML*, pp. 282–289 (2001).

[11] Liu, Y., Zhang, Y., Che, W., Liu, T. and Wu, F.: Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations, *EMNLP*, pp. 864–874 (2014).

[12] Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs), `http://www.chokkan.org/software/crfsuite/` (2007).

[13] 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，Vol. 28, No. 5, pp. 420–432 (2013).

表 7 P-CRF (BCCWJ+ins1del0)，P-CRF (BCCWJ+ins2del0) 間の単語分割悪化例

P-CRF (BCCWJ+ins1del0)	P-CRF (BCCWJ+ins2del0)
ダフト｜パンク｜は｜人気｜が｜あ｜た｜と｜ 思い ｜まし｜た｜。	ダフト｜パンク｜は｜人気｜が｜あ｜た｜と｜ 思｜い ｜まし｜た｜。
CO ｜は｜部屋｜を｜買い｜に｜近い｜所｜で｜探そう｜の｜で｜、｜彼女｜の｜代わり｜今日｜は｜私｜会社｜を｜当番｜に｜ 行き ｜まし｜た｜。	CO ｜は｜部屋｜を｜買い｜に｜近い｜所｜で｜探そう｜の｜で｜、｜彼女｜の｜代わり｜今日｜は｜私｜会社｜を｜当番｜に｜ 行｜き ｜まし｜た｜。
できる｜こと｜なら｜、｜あの｜頃｜に｜戻っ｜て｜、｜人生｜を｜やり直し｜て｜、｜夢｜を｜ 叶い ｜ます｜か｜。	できる｜こと｜なら｜、｜あの｜頃｜に｜戻っ｜て｜、｜人生｜を｜やり直し｜て｜、｜夢｜を｜ 叶｜い ｜ます｜か｜。

表 8 P-CRF (BCCWJ)，P-CRF (BCCWJ+ins0del1) 間の単語分割悪化例

P-CRF (BCCWJ)	P-CRF (BCCWJ+ins0del1)
で｜も｜、｜ 付き合う｜の ｜時間｜が｜長い｜に｜なっ｜たら｜、	で｜も｜、｜ 付き合うの ｜時間｜が｜長い｜に｜なっ｜たら｜、
明日｜は ｜、｜ベルリン｜に｜、｜3｜8｜度｜でしょう｜。	明日は ｜、｜ベルリン｜に｜、｜3｜8｜度｜でしょう｜。
ここ｜は｜ 寒い｜だ ｜から｜ジョギング｜でき｜ない｜。	ここ｜は｜ 寒いだ ｜から｜ジョギング｜でき｜ない｜。

表 9 ins0del1 中の学習者文と添削文のペア

学習者文	添削文
現在の世界は以前 の より、怖くなる一方です。	現在の世界は以前より、怖くなる一方です。
まあ、すぐに終われるものじゃない だ から、これからもゆっくり楽しみましょう。	まあ、すぐに終われるものじゃないから、これからもゆっくり楽しみましょう。
恐らく来週 に アメリカ帰国する	恐らく来週アメリカ帰国する

表 10 KyTea (BCCWJ+ins1del0)，P-CRF (BCCWJ+ins1del0) 間の単語分割改善例

KyTea (BCCWJ+ins1del0)	P-CRF (BCCWJ+ins1del0)
ダフト｜パンク｜は｜人気｜が｜ あた ｜と｜思い｜まし｜た｜。	ダフト｜パンク｜は｜人気｜が｜ あ｜た ｜と｜思い｜まし｜た｜。
待っ｜て｜ ほが ｜いい｜です｜ね｜?	待っ｜て｜ ほ｜が ｜いい｜です｜ね｜?

表 11 KyTea (高性能 SVM モデル)，P-CRF (BCCWJ+ins1del0) 間の単語分割改善例

KyTea (高性能 SVM モデル)	P-CRF (BCCWJ+ins1del0)
毎日｜毎日｜、｜日本｜語｜を｜練習｜し｜ ましょ｜う ｜！	毎日｜毎日｜、｜日本｜語｜を｜練習｜し｜ ましょう ｜！
ビール｜を｜飲め｜ば｜後｜課題｜が｜でき｜ない｜よう｜に｜なる｜ん｜ だろ｜う ｜と｜思っ｜て｜…	ビール｜を｜飲め｜ば｜後｜課題｜が｜でき｜ない｜よう｜に｜なる｜ん｜ だろう ｜と｜思っ｜て｜…

表 12 MeCab, P-CRF (BCCWJ+ins1del0) 間の単語分割改善例

MeCab	P-CRF (BCCWJ+ins1del0)
そんな｜景色｜は｜ トラ｜マ ｜を｜よく｜ みえ｜り ｜ます｜。	そんな｜景色｜は｜ トラマ ｜を｜よく｜ みえり ｜ます｜
おもしろい｜ブログ｜を｜書き｜たい｜です｜が｜なに｜も｜ がん｜が｜えら ｜ない｜。	おもしろい｜ブログ｜を｜書き｜たい｜です｜が｜なに｜も｜ がんがえ｜ら ｜ない｜。
今日｜、｜友達｜と｜一緒｜に｜ ケーき屋 ｜を｜通し｜た｜とき｜、｜後｜で｜ケーキ｜を｜買おう｜と｜友達｜が｜言っ｜た｜とき｜、｜突然｜、｜今日｜は｜友達｜の｜誕生｜日｜と｜さっぱり｜思い出し｜た｜。	今日｜、｜友達｜と｜一緒｜に｜ ケーき｜屋 ｜を｜通し｜た｜とき｜、｜後｜で｜ケーキ｜を｜買おう｜と｜友達｜が｜言っ｜た｜とき｜、｜突然｜、｜今日｜は｜友達｜の｜誕生｜日｜と｜さっぱり｜思い出し｜た｜。

表 13 MeCab, P-CRF (BCCWJ+ins1del0) 間の単語分割悪化例

MeCab	P-CRF (BCCWJ+ins1del0)
「｜ ほうき星 ｜」｜の｜譜｜が｜欲しい｜です｜。	「｜ ほうき｜星 ｜」｜の｜譜｜が｜欲しい｜です｜。
この｜まま｜ 捨て置く ｜わけ｜に｜は｜ゆか｜ぬ｜」｜と｜仰い｜まし｜た｜。	この｜まま｜ 捨て｜置く ｜わけ｜に｜は｜ゆかぬ｜」｜と｜仰い｜まし｜た｜。
十分｜配慮｜さ｜れ｜て｜い｜ない｜アンケート｜ で｜すい ｜ませ｜ん｜。	十｜分｜配慮｜さ｜れ｜て｜い｜ない｜アンケート｜ ですい ｜ませ｜ん｜。
赤｜信号｜の｜ 時｜道 ｜を｜渡っ｜て｜は｜いけ｜ませ｜ん｜よ｜。	赤｜信号｜の｜ 時道 ｜を｜渡っ｜て｜は｜いけ｜ませ｜ん｜よ｜。