CRF を用いた漢文の返り点推定

情報科学科　佐藤 綾花　　指導教員：山村 毅

1　はじめに

古典中国語である漢文は, 日本の学校教育において, 古典の授業の中でも取り扱われており, 日本人にとって馴染みの深いものである. ほとんどの日本人は, 漢文を読む際に, 返り点[*1]の指示に従い, 古典日本語に変換して読んでいる. これは, 本来の文法通りの漢文を読む知識がなくとも, 返り点の情報さえあれば翻訳できることを示している. よって, 返り点は漢文を理解するうえで重要な情報である. 先行研究 [1](2011) では, 返り点に着目し, 統計的推定手法を用いて, 白文[*2]に適切な返り点を推定する手法を提案した. そこでは, 漢字個々については, 72.83%, 漢文一文については, 26.83% の精度が得られている. 本研究では, この先行研究の反省を踏まえて, より正確に推定できる方法の開発を行う. 本研究で得られた成果は, 漢文研究の一助となり, また, 古典分野の自然言語処理におけるコーパス作成の支援にもなる.

2　提案手法

2.1　CRF を用いた返り点の推定

まず, 前処理として古典中国語形態素解析器 [2] を利用して, 全文章を単語に分解する. また, 本研究では返り点の推定を, 系列ラベリング問題として扱い, 各単語列に対して, 返り点ラベル列を, CRF（条件付き確率場）を用いて推定する. 推定に用いた素性は, 「単語そのもの」「大品詞」「品詞」「文末か否か, または読点の手前か」「文頭か否か」「文頭からの順番」「単語の長さ」「返読文字か否か」の 8 つである.

2.2　返り点の並びに対する修正

2.1 で推定した結果から, さらに返り点の規則に違反しているものを修正する. 修正には, CRF を用いた際に得られる周辺確率を利用し, 返り点の規則が正しく, かつ確率値の合計が大きくなるものを正解とする.

3　評価実験

3.1　実験方法

3 種類のテキストデータを用いて実験を行った. 1 つ目は, 先行研究でも用いた, 唐詩三百詳解上巻にある, 漢詩の漢文のデータ 410 文章であり, これを韻文テキストと呼ぶことにする. 2 つ目は, 改訂版高校学校古典漢文編にある, 物語の漢文のデータ 427 文章であり, これを散文テキストと呼ぶことにする. 3 つ目は, 単純に, 韻文テキストと散文テキストを混ぜたテキスト 837 文章であり, これを混合テキストと呼ぶことにする. これら 3 つのテキストを用いて, 2 で述べた方法で返り点を推定し, 精度を評価した. 評価には 10 分割の交差検定を用いた. また, CRF の実装には CRF++[3] を利用している. 結果を表 1 に示す.

3.2　考察

まず, 先行研究との比較をする. 先行研究では, 韻文テキストを用いて実験しており, 文章ごとの精度は 410 文章中 110 文章であった. したがって, 128 文章以上が正解していれば有意な差（有意水準 5% の適合度検定）があるといえる.

[*1] 返り点: レ点など, 変換の規則を明示的な形で漢字の傍に示したもの
[*2] 白文: 返り点のついていない漢文のこと

表 1　評価実験の結果 (文章)

	文章正解数	文章総数	精度 (%)
(1-a) 韻文テキスト	143	410	34.9
(2-a) 散文テキスト	119	427	27.9
(3-a) 混合テキスト	250	837	29.9
(1-b) 韻文テキスト（修正後）	217	410	52.9
(2-b) 散文テキスト（修正後）	146	427	34.2
(3-b) 混合テキスト（修正後）	354	837	42.3

本研究では返り点修正前で 143 文章, 修正後で 217 文章が正解であったので, 先行研究と比較して, 有意な差がある（精度が良い）といえる.

また, 表 2 は, 返り点の修正前後の結果を, 適合度の検定によって比較したものであるが, どのテキストにおいても, 有意な差を示すことができ, この修正が有効であることがわかる.

表 2　修正前後の正解数と有意差の比較（有意水準 5%）

	修正前	有意な差	修正後
(1-a) 韻文テキスト	143	163 以上	217
(2-a) 散文テキスト	119	138 以上	146
(3-a) 混合テキスト	250	277 以上	354

4　まとめ

本研究では, CRF を用いた返り点の推定手法を提案し, 先行研究よりも高い精度で推定することができることを示した. また, 返り点の規則に従い, 推定結果を修正する手法を提案し, さらに精度を向上させることができた.

先行研究と比較して精度は向上したものの, 特に, 散文テキストについては 34.2% とまだ改善の余地がある. 今後の課題として, まず, より有効な素性を探すことが挙げられる. 本研究で用いた素性は 8 つであったが, 例えば, 漢字のクラスタリングを行うなどして, さらに素性を追加することにより, 精度を向上させられる可能性がある.

また, 返り点の根本的な「飛ぶ-戻る」という性質に注目した, 新しい推定方法を考案することが挙げられる. 本研究では, 最大 10 種類の返り点ラベルを推定していたが, 返り点の種類で推定するのではなく, 返り点の性質を推定することによって, 推定するラベルの数が減り, 出現頻度の低い返り点ラベルも推定できるようになる可能性がある.

参考文献

[1] 佐藤綾花, 小林真也, 山村毅: ”漢文における返り点の統計的推定手法”, 電気関係学会東海支部連合大会講演論文集, P2-6, 静岡大学, 2013

[2] 守岡知彦: ”MeCab を用いた古典中国語形態素解析器の改良”, 情報処理学会研究報告, 2009-CH-84, pp1-5,2009

[3] Taku Kudo: ”CRF++: Yet Another CRF toolkit”, http://crfpp.googlecode.com/svn/trunk/doc/index.html