近代の歴史的資料を対象とした機械学習による文境界推定

白井良介† 松村雪桜† 小木曽智信‡ 小町守†
首都大学東京† 国立国語研究所‡

shirai-ryosuke@ed.tmu.ac.jp, matsumura-yukio@ed.tmu.ac.jp,
togiso@ninjal.ac.jp, komachi@tmu.ac.jp

1 はじめに

文境界推定はあらゆる自然言語処理の分野において必要不可欠となる要素技術である。形態素解析や固有表現抽出、係り受け解析などのタスクでは、文書ではなくそれぞれの文に対して解析を行うため、正しい文境界が定まっていることが前提になっている。

現代の日本語の書き言葉においては“。”やエクスクラメーションマーク、クエスチョンマークが手がかりとなっているため文境界の付与が容易である。その一方で、ウェブのテキストのように自由記述形式のものや、話し言葉の書き起こし、また**歴史的資料においては手がかりが曖昧であり、ルールベースで文境界を付与することが困難**な場合が存在する。特に、近代の歴史的資料に対して文境界を付与することは近代語の知識のある専門家の手に依らなければ難しく、膨大な量の資料の前に作業がなかなか進まないでいるのが現状である。現在、近代の歴史的資料に対しては専門家らによる人手のアノテーションが行われているが、まだアノテーションのなされていない膨大な量の資料が存在する。アノテーションは専門家らの知識を前提として成り立っており、ルールベースでの学習は困難である。

そこで、本研究では、近代の歴史的資料を対象に機械学習による文境界推定を行う。**ルールベース以上に複雑な素性を扱うことができる機械学習を用いた文境界推定を行う**ことで、膨大な量の資料に対して一次的なアノテーションを施すことができるということが本研究の貢献である。また、近代の歴史的資料を対象にした機械学習による文境界推定を行うのは本研究が初めてである。

本研究で文境界推定を行う資料は、1895 年（明治 28 年）から 1928 年（昭和 3 年）に博文館より発行された総合雑誌『太陽』を対象とし、データは『太陽コーパス』[4] の文語コアデータを用いた。“。”で文境界を付与するルールベースのものをベースラインとし、『太陽コーパス』のみを用いて学習したモデル、『太陽コーパス』に『太陽』と同時代の資料を加えて学習したモデルを用いて、文境界推定との異なり具合と、近代語への文境界推定の精度を確認した。機械学習の手法としては条件付き確率場（CRF）[1] を使用し、文境界推定の閾値を調整することで、適合率と再現率のトレードオフをコントロールすることができる。また、ベースライン（ルールベース）の適合率 95.0%・再現率 33.5%の精度から、適合率 86.1%・再現率 71.0%と適合率は少々下がってしまったが再現率を大幅に上げることができた。提案手法の文境界の閾値を適合率が 95.0%になる値に調整したところ、再現率が 41.9%となり、ベースラインと比較して同じ適合率でも高い再現率を出すことができた。

2 関連研究

現代の書き言葉を対象にした文境界推定は、いくつか研究が行われている。例えば、英語では文境界を表すピリオドと“Mr.”などのように文境界を表さないピリオドが存在するため、書き言葉に対する文境界推定を行う必要がある [2]。

日本語の文境界推定の研究として行われているのは、推定の対象として主にウェブテキストのような自由記述形式の書き言葉を対象としたもの [7] や話し言葉の書き起こしを対象としたもの [9] である。これらは日本語の書き言葉の文境界を表す“。”などの目印が必ずしも付与されていないため、文境界の推定が必要である。本研究の対象である近代の歴史的資料についても、文境界の手かがりとなるものが必ずしも存在しないため、ウェブテキストやスピーチの書き起こしと同様に文境界推定を行うことが必要である。

表 1: 近代の歴史的資料における 4 つの文パターン

パターンと例文（文境界を “|” で示す）	
現代語パターン:	“、” と “。” が現代語の書き言葉と同じように付与されている
例:	|一は歐羅巴の海岸線が甚だ複雑なる事にして、一は其上に位する國民の種類の甚だ夥多なる事なり。|
“。” パターン:	“。” を “、” の役割としても付与している
例:	|おや。二個貰ツたのか。|
“、” パターン:	“、” を “。” の役割としても付与している全 “、” パターン
	段落終わりのみ “。” を付与している例外パターンも存在する
例 1:	|記者曰、君は徳太郎と稱し、慶應三年十二月を以て江戸芝神明町に生る、|
例 2:	|豈に戒めざる可けんや、|豈に懼れざる可けんや。|（段落終）
“。”・“、” なしパターン:	そもそも “、” と “。” が付与されていない
例:	|請ふ其の昨年度の形勢を觀察せん|今昨年五月末日に於ける船舶の統計は左の如し|

表 2: 各近代語コーパスにおける文パターンの割合と統計情報

コーパス	現代語	“。”	“、”	“。”・“、”	短単位数	文書数	文数
『太陽コーパス』文語	30.8%	3 文のみ	35.7%	33.4%	71,850	33	3,686
『明六雑誌コーパス』	0.0%	0.0%	3.3%	96.7%	179,522	198	9,563
『国民之友コーパス』	11.0%	1 文のみ	21.8%	67.1%	32,154	24	1,479
『女性雑誌コーパス』文語	30.2%	2 文のみ	31.7%	38.7%	39,779	64	2,148

3 機械学習を用いた文境界推定

本研究では、文境界推定を文頭の形態素に対応する B ラベルと文頭でない形態素に対応する I ラベルを予測する BI ラベルの系列ラベリング問題としてとらえ、機械学習を行った。

3.1 文パターンの統一

近代の歴史的資料において文境界を推定することが困難な理由としては、4 つのパターンの文の記述が混在していることが挙げられる。表 1 に『太陽コーパス』の文語コアデータにおける各パターンの例文を示し、表 2 に今回実験に用いた各コーパスにおけるデータ内での割合と統計情報を示した。**現代語**パターン以外にも **“。”** パターンと **“、”** パターン、そして **“。”・“、” なし**パターンが存在し、後者の 3 パターンはルールベースで解析することができない。『太陽コーパス』以外のコーパスについては 4.1 節で詳しく述べる。

表 1 に示したように、近代語コーパスでは “、” と “。” が文書中に混在しており、またその役割も一様ではないために 4 番目の **“。”・“、” なし**パターンに統一し、“、” と “。” を全て取り除くこととした。

また、文中に出現する全角スペースを取り除いた。これは歴史的資料によく見られる決まり事である、‘‘皇室関係者の名前を記す際には敬意を表して該当する用語の前に空白を付する（闕字）” ということと、段落はじめの全角スペースに B ラベルが振られていた場合、次に出現する本来は文頭になるべき形態素に I ラベルが振られているということを考慮したためである。その場合、文頭の全角スペースに振られていた B ラベルについては次に出現する形態素に付与し直すことで対応した。

表 3: 素性テンプレート

N-gram	観測するトークン
uni-gram	$x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}$
bi-gram	$x_{t-2}x_{t-1}, x_{t-1}x_t, x_tx_{t+1}, x_{t+1}x_{t+2}$
tri-gram	$x_{t-2}x_{t-1}x_t, x_{t-1}x_tx_{t+1}, x_tx_{t+1}x_{t+2}$

3.2 素性テンプレート

素性テンプレートには近代文語 UniDic [5] で定義される素性のうち、1. 書字形出現形 (orth)、2. 品詞 (pos)、3. 活用形 (cForm)、4. 語彙素表記 (lemma) の 4 種類を用いた。

それぞれ、現在のトークンを x_t としたとき、現在のトークンと前後 2 トークンずつの uni-gram、bi-gram、tri-gram を利用する。詳しくは表 3 に示した。

表 4: 文境界推定 実験結果

手法	学習データ	適合率	再現率	F 値
ルールベース	−	95.0%	33.5%	49.5
CRF	『太陽』のみ	79.7%	66.5%	72.5
CRF	+3 コーパス	86.1%	71.0%	77.8

4 近代語に対する文境界推定実験

近代語の文境界推定において、コーパスの形態素に対して系列ラベリングを適用し、様々な学習データのパターンから『太陽コーパス』のコアデータのうち文語データにおける文境界推定の性能を比較した。形態素として近代文語 UniDic の短単位 [3] を用いた。評価には再現率、適合率、F 値を用いた。実装は CRF++ [1] を使用した。実験時のパラメータにはデフォルト値を用いた。

4.1 データ

実験対象の近代語資料として『太陽』の人手で修正が行われているコアデータを用いた。『太陽コーパス』には文語・口語の両データが存在するが、近代の資料には文語体で記述されたものが多いことを考慮して、より多くの資料に対して文境界を推定できるモデルを構築するために文語データのみを用いて 5 分割交差検証を行った。5 分割は全 35 文書からなる文語データをランダムに 7 文書ずつ抽出することにより行った。

また、学習データの不足を考慮して、追加の学習データとして『太陽コーパス』と同じく近代語コーパスである、『明六雑誌コーパス』[8]『国民之友コーパス』[8]『女性雑誌コーパス』[6] の 3 種を用いることにした。『女性雑誌コーパス』については、『太陽コーパス』と同様に文語・口語の両データが存在するため、文語データのみを用いた。それぞれのコーパスの総短単位数と総文数を表 2 に示した。

4.2 手法

ベースラインとして “。” を文境界と認定する手法を用いたルールベースの実験を行った。ルールベースの実験の際には、3.1 節で述べた “。”・“、” の除去は行なっていない。また、提案手法として前節で述べた系列ラベリングを用いて実験を行った。評価には B ラベル推定の適合率、再現率、F 値を用いた。また、各文書の最初のトークンが B ラベルであることは自明なので、該当するトークンは評価対象から外した。

[1] https://taku910.github.io/crfpp/

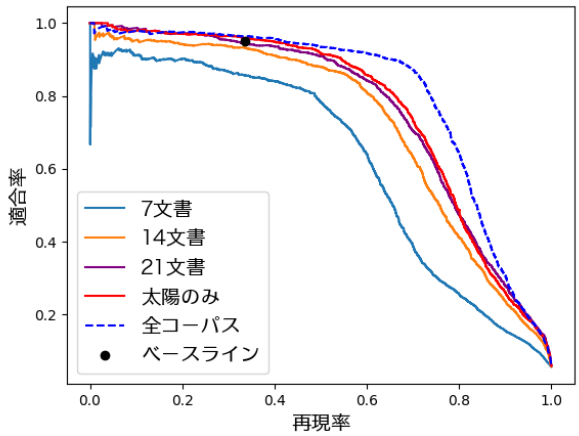

図 1: 適合率 - 再現率曲線

4.3 実験結果

表 4 に結果を示し、図 1 に適合率 - 再現率曲線を示した。実験の結果、ベースラインと比較して『太陽コーパス』のみを用いた実験で 3.9 ポイント高い F 値を得ることができ、『太陽コーパス』に 3 種のコーパスを追加した実験では 9.3 ポイント高い F 値を得ることができた。

5 考察・エラー分析

ルールベースを用いた実験結果の再現率が最も低く、近代の資料に対しては現代語と同じような手法では文境界推定のカバー率を上げるのが難しいことがわかる。近代語では動詞 “有り” 等のラ行変格活用の語が頻出するが、現代語であればともに “有る” となる終止形と連体形がそれぞれ “有り” と “有る” で異なる一方、終止形と連用形がともに “有り” で同形となる。そのため、文境界認定の上では現代語とは異なって、中止と終止の区別が付けづらいという特徴がある。

いずれの実験結果でも適合率に比べて再現率が低く、文境界があるべき場所を正しく検出するのは難しいという結果となった。

学習用データとして『太陽』のみを用いた実験結果と、『太陽』・『明六雑誌』・『国民之友』・『女性雑誌』を用いた実験結果では、後者の方が適合率・再現率・F 値のいずれにおいても高くなっている。『太陽コーパス』文語データのみを用いた時の学習曲線を図 1 に示す。この曲線が示すようにまだ伸びしろがあり、同じく近代の文語体で書かれているデータを追加することで再現率を上げることができた。

表 5: FN の頻出のエラー 5 件

間違えたトークン	全 FN に占める割合
と	9.0%
同	2.1%
其（其の）	1.7%
今	0.9%
是（此れ）	0.8%

もっとも精度が高い全コーパスを学習に用いた実験結果の中で生じた全エラー 1,486 個のうち、false negative が 1,133 個（76.2%）、false positive が 353 個（23.8%）であった。再現率を高くするために改善が必要な false negative（FN）の中から割合の高いものを表 5 に示す。その際、“其” は “其の” という連体詞と、‘‘其れ” という代名詞の両方で “其” という同じ表記であるため、連体詞の “其の” であることを括弧で示した。同様の理由で “是” は “此処” という代名詞と “此れ” という代名詞の区別を括弧で示した。

個別のトークンとして最も割合の高い、“と” は、‘‘夫れは君の意見に任せるⅠと言ひます” のように直前が文境界となり “と” が B ラベルになる場合と、“波蘭統監に任ずとⅠ”、“狩野氏と志筑氏” のように直前が文境界とならず “と” が I ラベルになる場合があり、‘‘任せる”、“任ず” のように終止形の後ろに “と” が出現する場合でも推定するラベルに異なりがある。“波蘭統監に任ずとⅠ” のような終止形で終わる文では、終わったあとの文末に “と” を終端記号のように付与していることが特徴である。“と” には格助詞、接続助詞、係助詞など様々な用例があり、品詞推定によってこれらの用例は区別できるが、品詞の同定にも曖昧性がある。2 番目に割合の高い “同” は、‘‘同世紀”、“同教授” のように前述のものと同じという意味で使用されていることが多いが、“同” ＋日付で小見出しとして B ラベルになるパターンが存在し、そこで揺れが生じたと考えられる。残りの頻出エラーである “其”、“今”、“是” については、はっきりとしたエラーの傾向が発見できず、出現回数の多い短単位であるため必然的にエラーの発生数が多くなってしまい、合計した結果、エラー頻出率の上位に入ってしまった可能性があると考えられる。

また、学習データでの出現回数が極めて少ない短単位で起こるエラーが 61%にもなる。そのため、同時代の学習データを増やし半教師あり学習によって再現率を改善していくことが期待される。

6 おわりに

本研究では、近代語の歴史的資料に対する CRF を用いた機械学習による文境界推定を行った。専門家によるアノテーションを待っている膨大な量のデータに付与することができる一次的な文境界としての活用が期待される。

今後の課題としては、今回の研究では “。”・“、” をすべて取り去って文境界推定を行ったが、“。”・“、” の両方、あるいは片方のみが付与されているものと付与されていないものに分け、付与されているものについては “。”・“、” でない箇所は文境界とならないという制約を設けることで、5 節で述べた “波蘭統監に任ずとⅠ” のような “。”・“、” が付与されていない箇所でのエラーを除くことができると考えられる。

付記

本研究は国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の研究成果の一部を報告したものである。

参考文献

[1] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pp. 282–289, 2001.

[2] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen SOLBERG Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING*, 2012.

[3] 近藤明日子. 近代文語 UniDic 短単位規定集 Ver.1.1. 2016.

[4] 国立国語研究所. 『太陽コーパス―雑誌「太陽」日本語データベース―』. 博文館新社, 2005.

[5] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, Vol. 20, No. 5, pp. 727–748, 2013.

[6] 田中牧郎. 『近代女性雑誌コーパス』の概要. 日本学術振興会科学研究費補助金研究成果報告書 基盤研究 (B)「20 世紀初期総合雑誌コーパス」の構築による確立期現代語の高精度な記述』pp.55-62, 2006.

[7] 福岡健太, 松本裕治. Support vector machines を用いた日本語書き言葉の文境界推定. 言語処理学会年次大会発表論文集, pp. 1221–1224, 2005.

[8] 近藤明日子. 『明六雑誌コーパス』『国民之友コーパス』の構築. 日本語の研究, Vol. 12, No. 4, pp. 167–174, 2016.

[9] 下岡和也, 内元清貴, 河原達也, 井佐原均. 日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化. 自然言語処理, Vol. 12, No. 3, pp. 3–17, 2005.