学修番号 17890520

修士論文

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

2019 年 2 月 22 日

首都大学東京大学院
システムデザイン研究科 情報通信システム学域

白井　良介

審査委員：

小町 守 准教授　（主指導教員）

石川 博 教授　（副指導教員）

片山 薫 准教授　（副指導教員）

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

修論要旨

本稿では、機械学習を用いて近代の歴史的資料に対して文境界を検出する手法を提案する。

文境界推定は多くの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。しかし、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しており、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いるため、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。ルールベースに対して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文境界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、日本語の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。

*首都大学東京大学院 システムデザイン研究科 情報通信システム学域 修士論文, 学修番号 17890520, 2019 年 2 月 22 日.

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』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と “。”・“、” で文境界を付与したルールベースのもの、2 つ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（Conditional Random Fields: CRF）と、文字単位の GRU（Gated Recurrent Unit）[1] を双方向に用いた Bi-GRU（Bi-directional GRU）を使用した。ベースライン（ルールベース）の適合率 94.34%・再現率 34.81%・F 値 50.85 ポイントの精度と比較して、CRF を用いた手法では適合率 83.75%・再現率 73.68%・F 値 78.40 ポイント、Bi-GRU を用いた手法では適合率 75.07%・再現率 62.01%・F 値 67.08 ポイントと F 値を大きく向上させることができた。

上記の実験に加えて、本研究の提案手法で文境界を付与することが具体的にどう役立つかということを確認するために、文境界推定によって得られた文境界を与えて、形態素解析の精度を比較した。『太陽コーパス』に 3 種のコーパスを追加した文境界推定実験で付与した文境界が、ルールベースと比べて 0.26 ポイント高い F 値を得ることができた。ブートストラップ検定を行ったところベースラインに対して統計的に有意（$p < 0.001$）であり、形態素解析の前処理としても役立つことを示した。また、実際の文境界修正作業を模したアノテーション支援実験も行った。結果として、文書に対してルールベースの文境界が付与されているものに比べて、提案手法による文境界が付与されているほうが文境界修正作業の時間を大きく短縮することができた。

本研究の貢献は以下である。

- 活字資料が多く残っている近代語の資料について、そのうちデジタルに翻刻がなされた膨大な量の資料に対して、文境界アノテーションを施すことで、専門家らによるアノテーションが付与される前段階として、文書に施される自動付与されたタグの精度を向上させることができる。
- 上記に加えて、専門家らによるアノテーション付与の手助けとなることでアノテーション時間の短縮ができる。

本稿の構成は以下のようになっている。第 1 章 では本研究全体の提案、貢献、概要を述べる。第 2 章 では文境界推定に関する関連研究について述べる。第 3 章 では機械学習を用いた文境界推定についての設定や使用する素性について述べる。第 4 章 では近代語に対する文境界推定実験について、データ、手法、実験結果、考察、エラー分析を述べる。第 5 章 では推定した文境界を用いた検証実験について、まず形態素解析の改善について、データ、手法、実験結果、考察、エラー分析を述べ、次にアノテーション支援実験の結果について述べる。第 6 章 では本研究のまとめ、今後の展望について述べる。

Machine Learning-based Sentence Boundary Detection for Modern Japanese Texts*

Shirai Ryosuke

Abstract

In this study, we propose a method to detect sentence boundaries for modern Japanese texts using machine learning. For modern Japanese texts, sentence boundaries are not explicitly marked so that human annotation is inevitable, but the annotation process is far from complete due to enormous number of materials. Therefore, we propose a method to detect sentence boundaries using machine learning. The main contribution of this study is that this method can support the annotation task as a primary annotation. We also show that the accuracy of morphological analysis can be improved by performing sentence boundary detection. Moreover, this is the first work to detect sentence boundaries by machine learning targeting modern historical materials.

*Master's Thesis, Department of Information and Communication Systems, Graduate School of System Design, Tokyo Metropolitan University, Student ID 17890520, February 22, 2019.

目次

図目次

第 1 章　はじめに

文境界推定は多くの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、Twitter や Facebook などのソーシャルメディアの投稿やマイクロブログ等のウェブ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 [2] 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。

国立国語研究所での近代の歴史的資料に対するアノテーションでは、生のテキストからはじまり、最終的に人手による修正を経た高精度な形態素解析済みのコーパスを作るまでを目処としており、可能であれば係り受け情報までを付与することを目指している。同研究所では、原本からデジタルデータへの書き起こしがされたものに対して句読点などを仮の文境界として形態素解析をしたノンコアデータ、それに対して専門家らによる人手の修正がなされたコアデータの 2 種類のデータを用意している。本研究の対象としている文境界推定は、書き起こしがされたデータに対する前処理の段階で使用され、ノンコアデータの代わりとして使用されることを目指している。このノンコアデータは国立国語研究所が公開しているコーパス検索アプリケーション『中納言』での検索にも用いられるため、国立国語研究所の研究員だけでなく、外部のコーパス言語学者にとっても役立つことが期待される。本研究の位置付けを図 1.1 に示した。専門家が翻刻作業をすることで原本である紙から電子の生テキストへの翻刻の際に同時に文境界を付与するということも可能であるが、紙から電子への翻刻は専門家でない作業者が担当しており、同時に行うことは

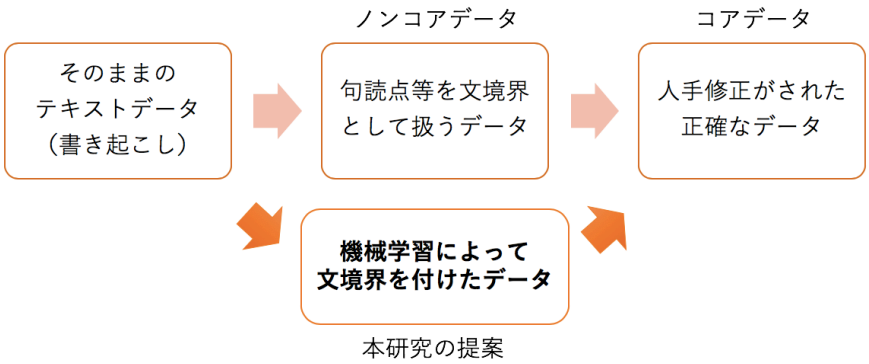

図 1.1: 本研究の位置付け

現実的には難しい。そのため文境界推定のタスクが果たす役割は大きいと言える。

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。ルールベースに対して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人手の修正が行われる前段階の一次的なアノテーションを改善することができるということが本研究の貢献である。また、日本の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』[3] 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と “。”・“、” で文境界を付与するルールベースの 2 つ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』と同時代の資料のコーパスである 3 種類の近代文語コーパス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、提案手法として短単位の素性テンプレートを用いた条件付き確率場（Conditional Random Fields: CRF）[4] と、文字単位の GRU（Gated Recurrent Unit）[1] を双方向に用いた Bi-GRU（Bi-directional GRU）を使用した。ベースライン（ルールベース）の適合率 94.34%・再現率 34.81%・F 値 50.85 ポイントの精度と比較して、『太陽』に 3 種類の近代文語コーパスを加えて学習し

た CRF を用いた手法では適合率 83.75%・再現率 73.68%・F 値 78.40 ポイント、Bi-GRU を用いた手法では適合率 75.07%・再現率 62.01%・F 値 67.08 ポイントと F 値を大きく向上させることができた。

上記の実験に加えて、本研究の提案手法で文境界を付与することが具体的にどう役立つかということを確認するために、文境界推定によって得られた文境界を与えて、形態素解析の精度を比較した。『太陽コーパス』に 3 種のコーパスを追加した文境界推定実験で付与した文境界が、ルールベースと比べて 0.02 ポイント高い F 値を得ることができた。ブートストラップ検定を行ったところベースラインに対して統計的に有意（$p < 0.001$）であり、形態素解析の前処理としても役立つことを示した。

また、実際の文境界修正作業を模したアノテーション支援実験も行った。結果として、文書に対してルールベースの文境界が付与されているものに比べて、『太陽』に 3 種類の近代文語コーパスを加えて学習した提案手法による文境界が付与されているほうが文境界修正作業の時間を大きく短縮することができた。

本稿の構成は以下のようになっている。第 1 章 では本研究全体の提案、貢献、概要を述べる。 第 2 章 では文境界推定に関する関連研究について述べる。第 3 章 では機械学習を用いた文境界推定についての設定や使用する素性について述べる。第 4 章 では近代語に対する文境界推定実験 について、データ、手法、実験結果、考察、エラー分析を述べる。第 5 章 では推定した文境界を用いた形態素解析の改善について、データ、手法、実験結果、考察、エラー分析と、推定した文境界を使った被験者実験の結果について記述する。第 6 章 では本研究のまとめ、今後の展望について述べる。

第 2 章　先行研究

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 [5]。

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 Twitter や Facebook などのソーシャルメディアの投稿やマイクロブログ等のウェブテキストのように自由記述形式の書き言葉を対象としたものや、話し言葉の書き起こしを対象としたものである。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。

文境界推定の方法には主に機械学習が用いられている。福岡ら [6] は Web 及びニュースグループから集めたテキストに対して SVM（Support Vector Machine）を用いて文境界推定を行なった。難波ら [7] は Twitter に投稿された Tweet を対象として、文境界推定を系列ラベリング問題として扱い、CRF を用いた文境界推定を行なった。CRF の素性には単語と品詞と文字種を使用して実験を行い、同時に文節境界推定と係り受け推定を行った。また日本語以外でも、Rudrapal ら [8] は英語やヒンドゥー語で書かれた Twitter の Tweet や Facebook のメッセージを対象として、CRF、ナイーブベイズ、SVM を用いて文境界推定を行った。話し言葉の書き起こしでは、下岡ら [9] は日本語話し言葉コーパス（CSJ）[10] を対象にして SVM を用いた文境界推定を行った。文境界推定をテキストチャンキングの問題として扱い、テキストチャンカとして SVM に基づく YamCha [11] を用いた。これらの研究のように、ウェブテキストやスピーチの書き起こしではルールに基づく処理が困難なため、機械学習による文境界推定が行われている。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、同様に文境界推定を行うことが必要である。

近年の機械学習のスタンダードであるニューラルネットワークを用いた研究では、Straka ら [12] が開発している UDPipe * というソフトウェアがある。これは文字単位の文字単位の GRU（Gated Recurrent Unit）[1] を双方向に用いた Bi-GRU

*http://ufal.mff.cuni.cz/udpipe

（Bi-directional GRU）を系列ラベリング手法として採用し、生テキストに対して単語分割、タグ付け、係り受けまでの解析をサポートしており、単語分割が行われるのと同時に文境界推定も行われている。本研究では上記の関連研究でも使用されている CRF と Bi-GRU の両手法を用いた文境界推定を行った。

また本研究と同様に、生テキストを対象としている CoNLL 2018 Shared Task[†] [13] では、生テキストから係り受けまでの解析を共通タスクとしており、古代ギリシア語、ゴート語、ラテン語、古代教会スラヴ語、古フランス語など本研究と同様に過去へ遡った言語が対象言語として含まれているが、日本語の近代の歴史的資料を対象として文境界推定を行うのは本研究が初めてである。

形態素解析の前処理の研究としては、統計的機械学習を用いて生の歴史的資料に対して濁点を付与する前処理を施す研究 [14] が行われている。近代の歴史的資料を対象として実験を行なっている点は共通しているが、両方は独立した手法であるため、組み合わせて使用することができる。

[†]http://universaldependencies.org/conll18/

第 3 章　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、人手でアノテーションされたデータを用いて CRF と Bi-GRU による機械学習を行うことで、文境界を自動で付与する手法を提案する。

3.1　4 つの文パターン

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 3.1 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 3.2 に今回実験に用いた各コーパスにおける文パターンの割合と統計情報を示した。“。”・“、”**混合**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。

3.2　CRF

機械学習の手法として CRF を用いる。実装には CRF++ *を使用した。素性には近代文語 UniDic [15]†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 の素性を利用する。詳しくは表 3.3 に示した。

*https://taku910.github.io/crfpp/

†近代文語 UniDic には『太陽コーパス』、『明六雑誌コーパス』、『国民之友コーパス』、『女性雑誌コーパス』に加えて、1 つのコーパスとしては扱われていない多くの近代論説文の語彙が収録されている。

表 3.1: 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）	
“。”・“、” 混合パターン:	“、” 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全 “。” パターン
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 3.2: 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	混合	“。”	“、”	“。”・“、” なし	短単位数	文書数	文数
『太陽』文語	30.8%	3 文のみ	35.7%	33.4%	71,850	33	3,686
『明六雑誌』	0.0%	0.0%	3.3%	96.7%	179,522	198	9,563
『国民之友』	11.0%	1 文のみ	21.8%	67.1%	32,154	24	1,479
『女性雑誌』文語	30.2%	2 文のみ	31.7%	38.7%	39,779	64	2,148

表 3.3: 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	$x_{t-2}x_{t-1}x_t$, $x_{t-1}x_tx_{t+1}$, $x_tx_{t+1}x_{t+2}$

3.3 Bi-GRU

機械学習のもう 1 の手法として Bi-GRU を用いる。文字単位の GRU を双方向に用いた Bi-GRU を実装している UDPipe というソフトウェアを使用した。UDPipe は CoNLL-U フォーマットのコーパスからモデルを構築し、解析結果を出力するソフトウェアである。構築したモデルを用いて、生テキストに対して単語分割、タグ付け、係り受けまでの解析をサポートしており、この単語分割が行われるのと同時に文境界推定も行う。

第 4 章　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 [16] を用いた。評価には B ラベル推定の再現率、適合率、F 値を用いた。CRF の実装には CRF++ を使用した。実験時のパラメータにはツールのデフォルト値を用いた。文字単位の Bi-GRU による文境界推定の実装には UDPipe の単語分割機能により出力される文境界を使用した。実験時のパラメータは予備実験の結果より dimension を 64 に、segment size を 200 に変更し、その他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』は当時もっともよく読まれた総合雑誌であり、政治・経済・世界情勢から科学・思想、文学作品までの様々な記事ジャンルが揃っている。『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 33 文書からなる文語データをランダムに 7 文書または 6 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』[2]、『国民之友コーパス』[2]、『女性雑誌コーパス』[17] 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 3.2 に示した。

本研究の文境界推定は生テキストに対して用いることを想定しているため、CRF を用いた実験では人手で修正された形態論情報が付与されているコアデータではなく、自動解析結果であるノンコアデータに相当するものをテストデータとして使用

表 4.1: 文境界推定 実験結果

文境界推定の手法名	単語分割手法	文分割手法	学習データ	適合率	再現率	F 値
“。” ルール	MeCab	“。”	−	94.34%	34.81%	50.85
“。”・“、” ルール	MeCab	“。”・“、”	−	42.92%	61.89%	50.67
UDPipe	UDPipe	UDPipe	『太陽』のみ	72.90%	63.94%	68.13
UDPipe	UDPipe	UDPipe	『太陽』+ 3 コーパス	75.41%	66.68%	70.82
“。”・“、” ルール + CRF	MeCab	CRF	『太陽』のみ	95.00%	34.51%	50.63
“。”・“、” ルール + CRF	MeCab	CRF	『太陽』+ 3 コーパス	82.87%	73.76%	78.05
UDPipe + CRF	UDPipe	CRF	『太陽』のみ	95.00%	34.51%	50.63
UDPipe + CRF	UDPipe	CRF	『太陽』+ 3 コーパス	82.26%	74.65%	78.38

する必要がある。そこで、2 つの疑似的なノンコアデータを作成した。1 つは “。”・“、” を文境界として文分割を行い、MeCab を使用して形態素情報を付与したもので、もう 1 つは UDPipe の単語分割によって得られた文境界で文分割を行い、同じく MeCab の部分的解析機能を使って単語境界以外を推定したものである。どちらも辞書には近代文語 UniDic を使用した。

4.2 手法

ベースラインとして “。” を文境界とする 1 つめのルールベース手法 (**“。” ルール**) と、“。”・“、” を文境界とする 2 つめのルールベース手法 (**“。”・“、” ルール**) 、UDPipe による文字単位の Bi-GRU を使用した手法 (UDPipe) を用意した。提案手法として **“。”・“、” ルール**の形態素解析結果を使った CRF と、UDPipe の形態素解析結果を使った CRF 実験を行った。機械学習を用いた手法では文書を入力とした。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。

4.3 実験結果

表 4.1 に適合率・再現率・F 値を示した。実験の結果、ベースラインのうち精度がより高かった **“。” ルール**手法と比較して『太陽コーパス』に 3 種のコーパスを追加した **“。” ルール** + CRF 手法の実験では 27.20 ポイント高い F 値を得ること

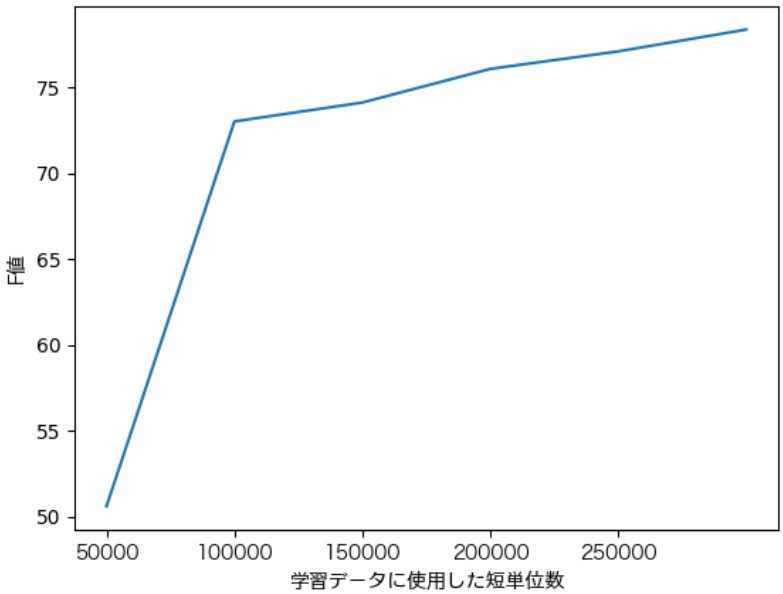

図 4.1: UDPipe+CRF モデルにまず『太陽』のデータを加え、順次 3 コーパスを加えていった場合の文境界推定の学習曲線

表 4.2: BCCWJ ルールベース適用結果

文書	適合率	再現率	F 値
OC（知恵袋）	81.06%	86.39%	83.64
OW（白書）	97.69%	63.18%	76.74
OY（ブログ）	80.98%	60.63%	69.35
PB（書籍）	97.64%	87.29%	92.18
PM（雑誌）	97.62%	73.91%	84.12
PN（新聞）	99.56%	71.99%	83.56

ができ、『太陽コーパス』に 3 種のコーパスを追加した UDPipe + CRF 手法の実験では 27.53 ポイント高い F 値を得ることができた。

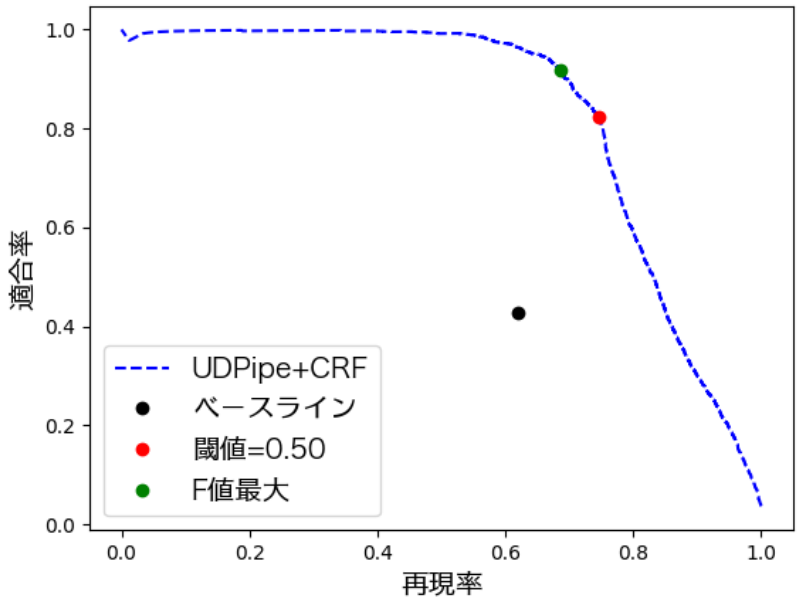

図 4.2: UDPipe+CRF モデルの PR 曲線

表 4.3: FN の頻出のエラー 上位 5 件

間違えたトークン	全 FN に占める割合
と	8.41%
全角空白スペース	3.88%
◎	2.04%
其	1.51%
今	1.05%

4.4 考察・エラー分析

“。”・“、” ルール手法を除いた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傾向があることがわかる結果となった。特に、“。” ルール手法の再現率は 2 番目に低く、近代の資

料に対しては現代語と同じような “。” を文境界として扱う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。比較対象の現代語の例として、現代日本語書き言葉均衡コーパス（BCCWJ）[18] のコアデータに対して “。”・エクスクラメーションマーク・クエスチョンマークを文境界として扱った際の B ラベル推定の再現率・適合率・F 値を表 4.2 に示した。現代語ではいずれのジャンルにおいても再現率 6 割以上になっているが、近代の資料に対しては、機械学習を用いなければ適合率を高く保ったまま再現率を上げることができない。その他の特徴として、近代文語では動詞 “あり” 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 “ある” となる終止形と連体形がそれぞれ “あり” と “ある” で異なる一方、終止形と連用形がともに “あり” 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、終止と中止の区別が付けづらいという特徴がある。

学習用データとして『太陽』のみを用いた手法と、『明六雑誌』·『国民之友』·『女性雑誌』それぞれのコーパスを追加して用いた手法では、後者の方が再現率・F 値が高くなっている。同じ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ることが確認された。『太陽』のみを学習に用いた場合では、**“。” ルール**手法とほとんど変わらない精度であった。また、**“。”·“、” ルール**手法と『太陽』のみで学習した UDPipe では後者のほうが F 値が 20.15 ポイント高いにも関わらず、『太陽』のみを学習に用いた CRF のモデルで文境界推定を行うと両者の精度に違いは現れなかった。このことから、前処理も精度に影響を与えるが、CRF のモデルによる影響のほうがより大きいことが確認できる。『太陽』にそれぞれのコーパスの短単位を追加していった時の学習曲線を図 4.1 に示す。この曲線が示すように、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 F 値を上げることができる。

図 4.2 に一番精度が高かった UDPipe+CRF 手法の PR 曲線を示す。適合率を “。”·“、” ルール手法と同じ 42.92% になるように調整した（閾値=0.02）ところ、UDPipe+CRF 手法の再現率は 85.11% となり、“。”·“、” ルール手法の再現率を 23.22% 上回った。また、F 値が最大（閾値=0.89）となるように閾値を調整しなくても、デフォルトの閾値でもルール手法より大きく適合率と再現率が向上していることからも、UDPipe+CRF 手法 が “。”·“、” ルール手法に対して精度的に優れていることが確認できる。なお、閾値には CRF の解析の周辺確率を用いた。

もっとも精度が高い『太陽コーパス』に 3 種のコーパスを追加した UDPipe + CRF 手法の実験結果の中で生じた全エラー 1,522 個のうち、false negative が 927 個（60.91%）、false positive が 595 個（39.09%）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 4.3 に示す。エラーについて考察を述べる際に、“|” で文境界を表す。

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 | 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 | ”、“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 | ” 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ため、品詞を素性に使用していたにも関わらずエラーが多くなってしまったと考えられる。2 番目に割合の高い “全角空白スペース” は、段落始めに頻出する形態素であるが、1 つで出現することもあれば数個続いて出現することもあり、小見出しでは文末に付与されることもあるため、B ラベルと I ラベルが混在しやすい特徴がある。また、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の影響もあり、識別が困難であったと考えられる。3 番目に割合の高い “◎” は、主に文書中の小見出しのような文に付与されている記号であるが、今回は最も頻出のエラーの “と” の直後や 2 番目に頻出の “全角空白スペース” の前後に出現する回数が多く、互いに影響しあって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” 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、エラーに出現する回数が 3 回以下と極めて少ないエラーが 34.30% にもなる。図 1 から分かるように、微量ではあるが学習データの増加が精度向上に効果的であることが示されているため、同時代の学習データを増やして改善していくことが期待される。

第 5 章　推定した文境界を用いた検証実験

本研究の提案手法で文境界を付与することが具体的にどのように役立つかということを確認するために、前項で推定した文境界を与えて、形態素解析の精度の比較を行った。加えて、アノテーション支援実験により作業効率の向上を確かめた。

5.1　形態素解析の精度比較実験

第 4 章で推定した文境界を与えて、形態素解析の精度の比較実験を行った。

5.1.1　データ

正解データには『太陽コーパス』文語コアデータの形態素情報を用い、前章で推定した 3 種類の文境界推定手法による形態素解析結果の比較実験を行った。

現在、ノンコアデータには “。” ・“、” ルールベースによる **“。” ・“、” ルール**手法の文境界が付与されているため、ルールベースには “。” ルールベースの **“。” ルール**ではなく **“。” ・“、” ルール**を使用した。それに加えて **“。” ・“、” ルール** + CRF 手法で付与した文境界、UDPipe + CRF 手法で付与した文境界、以上の 3 種類である。

5.1.2　手法

形態素解析には MeCab [19] を用いた。辞書には近代文語 UniDic の学習に使用しているコーパス群から、文境界推定実験時の 5 分割交差検証で用いた『太陽コーパス』のデータについて、近代文語 UniDic の収録語彙はデフォルトのまま、学習に使用しているコーパス群から、文境界推定実験時の 5 分割交差検証で用いた 5 分割した『太陽コーパス』のデータについて、それぞれ解析対象とするデータのみを除きパラメータ推定し直したものを使用した。つまり、5 分割されたそれぞれのデータに対して解析用の辞書を 5 つ作成した。

表 5.1: 形態素解析 実験結果

文境界推定の手法名	単語分割手法	文分割手法	学習データ	適合率	再現率	F 値
“。”・“、” ルール	MeCab	“。”・“、”	−	94.85%	94.88%	94.86
“。”・“、” ルール + CRF	MeCab	CRF	『太陽』+ 3 コーパス	*94.88%	*94.89%	*94.88
UDPipe + CRF	UDPipe	CRF	『太陽』+ 3 コーパス	*94.87%	*94.88%	*94.88

表 5.2: 形態素解析エラー上位 10 件と手法ごと出現数（品詞＋活用形＋語彙素）

正解	解析誤り	“。”・“、” ルール	UDPipe +CRF	“。”・“、” ルール **+CRF**
助動詞 + 終止形-一般 + ず	**助動詞 + 連用形-一般 + ず**	**80**	**70**	**80**
助動詞 + 連用形-ニ + なり-断定	助詞-格助詞 +*+ に	59	60	60
名詞-普通名詞-一般 +*+ 者	名詞-普通名詞-サ変可能 +*+ 物	55	55	55
助詞-格助詞 +*+ に	助動詞 + 連用形-ニ + なり-断定	51	52	52
動詞-非自立可能 + 終止形-一般 + 有る	**動詞-非自立可能 + 連用形-一般 + 有る**	**35**	**29**	**35**
助詞-接続助詞 +*+ も	助詞-係助詞 +*+ も	35	35	35
副詞 +*+ 又	接続詞 +*+ 又	28	28	28
動詞-一般 + 連用形-一般 + つく	動詞-非自立可能 + 連用形-一般 + 付く	26	26	26
助動詞 + 連用形-ニ + だ	助動詞 + 連用形-ニ + なり-断定	22	22	22
接頭辞 +*+ 低	形容詞-一般 + 語幹-一般 + 低い	21	21	21

表 5.3: 助動詞 “ず” ・動詞 “あり” 活用表

	未然形	連用形	終止形	連体形	已然形	命令形
助動詞 “ず”	ず（ざら）	ず（ざり）	ず	ぬ（ざる）	ね（ざれ）	ざれ
動詞 “あり”	あら	あり	あり	ある	あれ	あれ

また、比較実験のツールには MevAL* を使用した。

5.1.3 実験結果

表 5.1 に形態素解析実験の適合率・再現率・F 値を示す。また、ブートストラップ検定を行い統計的な有意であるかということも確認した。“*” はベースラインに対して統計的に有意（$p < 0.001$）であることを示している。実験の結果、ベースラインの **“。”・“、” ルール**手法と比較して UDPipe + CRF による文境界推定実験で付与した文境界と **“。”・“、” ルール** + CRF による文境界推定実験で付与した文境界がそれぞれ 0.02 ポイント高い F 値を得ることができた。また、前述のようにどちらの手法もベースラインに対して統計的に有意であるため、形態素解析の前処理としても役立つことを示した。

*https://teru-oka-1933.github.io/meval/

5.1.4　考察・エラー分析

エラー分析には形態素解析結果の中から、品詞 (pos)、活用形 (cForm)、語彙素表記 (lemma) を使用した。表 5.2 に “。”・“、” ルール手法、“。”・“、” ルール + CRF 手法、UDPipe + CRF 手法で付与した文境界で形態素解析を行った実験のエラー上位 10 件と、それらの出現数を示した。活用形が存在しないものについては “*” で表記した。終止形になるべき形態素が連用形と誤検出されているエラーについて、UDPipe + CRF 手法では出現数が減っていることが確認できる。“。”・“、” ルール + CRF 手法では “。”・“、” ルールと比べて精度は向上しているものの、頻出のエラーについては同程度しか対応できていないことがわかった。

近代語の中で特に活用を間違いやすいものとして、表 5.2 の上から 1 番目と 7 番目に見られる助動詞 “ず” と動詞 “あり” がある。表 5.3 に活用表を示す。括弧内は助動詞に接続する際の特殊な活用である。どちらも連用形と終止形が同じ形をしており、前後の形態素情報が判別の重要な手がかりとなる。提案手法により文境界が決まると EOS が分かるので、後ろの形態素情報が EOS であると分かることが終止形である手かがりとなるため、連体形と終止形の区別ができるようになる。そのため UDPipe + CRF 手法において形態素解析精度を向上させたと考えられる。4 章にて “。”・“、” ルール手法の精度よりも UDPipe 手法の精度が高いという実験結果が示されているが、その結果が形態素解析の精度改善の度合いにも現れていると考えられる。

5.2　アノテーション支援実験

“。”・“、” ルール手法で文境界が付与されたノンコアデータと提案手法で文境界を付与したデータとで、どの程度作業効率に影響を及ぼすかということを確かめるためにアノテーション支援実験を行った。このノンコアデータが実際にアノテーションの前処理として使われている形式である。

表 5.4: アノテーション支援実験に使用するデータ

データ名	“。”・“、”の有無	短単位数
A	なし	514
B	なし	582
C	あり	521
D	あり	529

表 5.5: 各実験協力者が文境界を修正したデータと順序

実験協力者	データと順序
W	Ar → Bm → Cr → Dm
X	Ar → Bm → Cr → Dm
Y	Am → Br → Cm → Dr
Z	Am → Br → Cm → Dr

5.2.1 データ

表 5.4 にそれぞれのデータの詳細を示す。“。”・“、”の有無については、出現しないデータと出現するデータを 2 つずつ用いることにした。本実験では、『太陽コーパス』文語コアデータの異なる文書から抽出した 4 つのデータ（A〜D とする）を用いる。

5.2.2 実験計画

4 人の協力者に **“。”・“、” ルール**手法もしくは UDPipe + CRF 手法で文境界が付与された 4 つの文書を与えて、それぞれの文書に対して文境界の修正を行わせた。表 5.5 に 4 人がそれぞれどのデータを対象に実験を行ったか示す。データ名の後ろについている “r” は rule の頭文字をとったもので、**“。”・“、” ルール**手法で文境界を付与していることを表し、“m” は machine learning の頭文字をとったもので、UDPipe + CRF 手法で文境界を付与していることを表す。また、同時に文

表 5.6: 各データごとの文境界修正にかかった時間

実験協力者	Ar	Bm	Cr	Dm
W	4 分 41 秒	1 分 49 秒	2 分 26 秒	1 分 11 秒
X	6 分 41 秒	4 分 50 秒	4 分 04 秒	3 分 20 秒
合計	11 分 22 秒	6 分 39 秒	6 分 30 秒	4 分 33 秒
実験協力者	Am	Br	Cm	Dr
Y	4 分 41 秒	7 分 08 秒	4 分 02 秒	3 分 22 秒
Z	5 分 36 秒	5 分 35 秒	3 分 51 秒	3 分 00 秒
合計	10 分 17 秒	12 分 43 秒	7 分 53 秒	6 分 22 秒

表 5.7: 提案手法–ルールベース間で短縮できた時間

	A	B	C	D	合計
時間差	–1 分 05 秒	–6 分 04 秒	+1 分 28 秒	–1 分 09 秒	–6 分 50 秒

境界の修正にかかる時間も計測した。計測時間から、どの程度アノテーションを支援することができているかを確認する。

5.2.3 実験結果

各データごとの文境界修正にかかった時間を表 5.6 に示す。また、各データごとの提案手法–ルールベース間で短縮できた時間を表 5.7 に示した。“+” はルールベースと比べて多くかかった時間、“–” はルールベースと比べて短縮できた時間を表す。

5.2.4 考察

データ全体として、提案手法による文境界が付与されているデータのほうがルールベースのものに比べて合計して 6 分 50 秒の時間を短縮することができた。また、提案手法を用いた結果をデータ別に見ていくと、“。”・“、” がない A・B の両方

で短縮に成功し、“。”・“、”がある C・D では、C においては短縮できなかったが、D では短縮に成功している。このことから、本研究による文境界の付与は特に“。”・“、”がない文書に対して大きく役立つことが確認できた。

第 6 章　おわりに

本研究では、近代語の歴史的資料に対する CRF と Bi-GRU 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。エラー分析の結果、出現する回数が 3 回以下と極めて少ないエラーが 33.9% にもなるため、同時代の学習データを増やして改善していくことが期待される。現在、国立国語研究所では近代語のコーパスとして『教科書コーパス』[20]、『東洋学芸雑誌コーパス』[21] の構築が行われている。それらのコーパスが完成し、学習に使用できるデータが増えることでエラーの改善に繋げることができると考えられる。一般的にニューラルネットワークではデータ量が肝要だと言われており、データが増えることで提案手法の形態素情報を作る段階に使用している UDPipe の精度がさらに向上することが期待され、そこから高精度な形態素情報を用いた CRF による文境界推定を行うことができると考えられる。

また、本研究の提案手法で文境界を付与することが具体的にどう役立つかということを確認するために、文境界推定によって得られた文境界を与えて、形態素解析の精度を比較した。提案手法である、『太陽』＋ 3 コーパスの学習データから UDPipe で文分割をした文の形態素情報を用いた CRF による文境界推定実験が、ルールベース手法で付与した文境界と比べて 0.02 ポイント高い F 値を得ることができた。また、形態素解析の前処理としても役立つことを示した。

加えて、アノテーション支援実験を行い、本研究による文境界付与が実際に人手で文境界を修正する際にも時間の短縮に繋がることを示した。

研究の将来性としては、近代語には活字史料が多く残されており、現在、雑誌や新聞データを中心にデジタル化が進んでいる [22] [23]。それらのデジタルデータを扱おうとする際に、本研究を有効に活用できるのではないかと考えられる。

発表リスト

1. 白井良介, 松村雪桜, 小木曽智信, 小町守. **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.** 言語処理学会第 24 回年次大会, pp.1023-1026. March 15, 2018.

謝辞

本論文の執筆において、指導教員の小町守准教授には大変お世話になりました。文学部の出身で外部からの進学である私に自然言語処理を学ぶ機会をくださいましたこと、感謝いたします。

また、本論文の副査を引き受けてくださいました石川博教授、片山薫准教授に感謝いたします。

加えて、人間文化研究機構 国立国語研究所の小木曽智信教授に感謝いたします。研究テーマの相談に乗っていただくだけでなく、必要な資料を提供していただきました。アノテーション支援実験に協力してくださいました国立国語研究所の研究員の皆様にも感謝申し上げます。

最後に、研究室の同期・先輩・後輩に感謝いたします。皆さんには研究のアドバイスをたくさんいただきましたし、情報系としての基礎的な知識が欠けている私に対しても懇切丁寧に解説をしてくださいました。息抜きのテレビゲームやボードゲーム、スポーツをしたことも記憶に残っています。同じ研究室で学べたことを嬉しく思います。

参考文献

[1] K. Cho, B. vanMerrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” Proceedings of SSST, pp.103–111, 2014.

[2] 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.

[3] 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.

[4] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” Proceedings of ICML, pp.282–289, 2001.

[5] J. Read, R. Dridan, S. Oepen, and L.J.S. Solberg, “Sentence Boundary Detection: A Long Solved Problem?,” Proceedings of COLING, pp.985–994, 2012.

[6] 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.

[7] 難波悟史，門内健太，但馬康宏，菊井玄一郎，“マイクロブログに対する文境界推定および係り受け解析，” 言語処理学会年次大会発表論文集，pp.107–111，2015.

[8] D. Rudrapal, A. Jamatia, K. Chakma, A. Das, and B. Gamback, “Sentence Boundary Detection for Social Media Text,” Proceedings of ICON, pp.254–260, 2015.

[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.

[10] 古井貞熙，前川喜久雄，井佐原均，“科学技術振興調整開放的融合研究推進精度－大規模コーパスに基づく『話し言葉工学』の構築－，” 日本音響学会誌，第56 巻，pp.752–755，2000.

[11] T. Kudo and Y. Matsumoto, “Chunking with Support Vector Machines,”

Proceedings of NAACL, pp.192–199, 2001.

[12] M. Straka, J. Hajic, and J. Straková, “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing,” Proceedings of LREC, pp.4290–4297, 2016.

[13] D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, “CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,” Proceedings of CoNLL, pp.1–21, 2018.

[14] 岡　照晃, 小町　守, 小木曽智信, 松本裕治, “統計的機械学習を用いた歴史的資料への濁点付与の自動化,” 情報処理学会論文誌, vol.54, no.4, pp.1641–1654, 2013.

[15] 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.

[16] 国立国語研究所コーパス開発センター（近藤明日子），“近代文語 UniDic 短単位規定集 Ver.1.1,” 2016.

[17] 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.

[18] 国立国語研究所コーパス開発センター，“『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版,” 国立国語研究所コーパス開発センター，2011.

[19] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proceedings of EMNLP, pp.230–237, 2004.

[20] 服部紀子，間淵洋子，近藤明日子，小木曽智信，“国定教科書のコーパス構築と公開,” 日本語学会 2018 年度秋季大会予稿集，pp.582–585，2018.

[21] 南雲千香子，近藤明日子，“『東洋学芸雑誌』コーパスの構築,” 通時コーパス活用班合同研究集会，2017.

[22] 美馬秀樹，丹治　信，増田勝也，太田　晋，“近代文献のデジタルアーカイブ化とテキストマイニング—岩波書店「思想」を題材に,” 研究報告人文科学とコンピュータ（CH），vol.2012，no.4，pp.1–8，2012.

[23] 間淵洋子，“明治・大正期『読売新聞』コーパスの構築と課題,” 言語処理学会

第 24 回年次大会 発表論文集，pp.500–503，2018.