Bi-LSTMを用いた中古日本語の文境界推定

鈴木 理紗[†], 川上 玲[‡,§], カラーヌワット タリン[§§,#], 北本 朝展[§§,#], 中澤 敏明[†], 苗村 健[†]

† 東京大学, ‡ 東京工業大学, § デンソーアイティーラボラトリ,
§§ ROIS-DS 人文学オープンデータ共同利用センター, # 国立情報学研究所

日本語概要

古典籍・古文書の可読性を向上できれば，文学，歴史，文化から災害記録など多くの研究を加速できる．このため，機械による自動翻刻への期待がある．文字認識やかな漢字変換など処理は様々にあるが，本稿ではその中の文境界推定に取り組む．形態素を入力とし，また音声認識における現代語での文境界推定で高い性能を誇る Bi-LSTM を用いて，中古日本語の文境界を推定するモデルを構築した．平安時代の文献からなるコーパスに適用し，PR 曲線の AUC で 0.894 と高精度な結果を得た．また，1 名の専門家からのフィードバックでも高評価を得た．

Sentence Boundary Estimation of Ancient Japanese Using Bi-LSTM

Lisa Suzuki[†], Rei Kawakami[‡,§], Tarin Clanuwat[§§,#],
Asanobu Kitamoto[§§,#], Toshiaki Nakazawa[†], Takeshi Naemura[†]

†The University of Tokyo, ‡Tokyo Institute of Technology, §Denso IT Laboratory,
§§ ROIS-DS Center for Open Data in the Humanities, # National Institute of Informatics

English Abstract

To improve the readability of ancient Japanese books and documents, processes such as old character recognition, punctuation, and Hiragana-Kanji conversion are required. The automation of these processes will accelerate many research areas, including literature, historical and cultural analysis, and disaster records. In this paper, we focus on sentence boundary estimation. We develop a model for estimating sentence boundaries in ancient Japanese using Bi-LSTM, which has a high performance of sentence boundary estimation in modern natural language processing for speech recognition. When applied to a corpus consisting of literature from the Heian period, the AUC of the PR curve achieved 0.894. The model was also highly evaluated by an expert.

1. はじめに

古典籍・古文書は現代の人間にとって可読性が低い．しかし源氏物語などの文学研究から，歴史や文化，地震などの研究にわたって，これらの読解を必要とする場面は多く存在する．そこで通常，翻刻を行い，可読性を向上させる．翻刻には専門知識が必要であるが，残存する文献数約 10 億冊に対し専門家は数千名にとどまるため，機械による自動翻刻の開発が期待される [19], [20].

翻刻は，図 1 に示すように，当時の文字（くずし字）を現代の文字へ変換する「翻字」と，変換後のテキストを可能な限り内容を変えずに修正し，可読性を高める「校訂」に分かれる．翻字には光学的文字認識 [4], [10], [11] や，市民が翻字に参加するクラウドソーシング [18] といった取り組みが存在するため，本稿は校訂の自動化に着目する．校訂には，文境界の挿入， 発話箇所への鉤括弧の挿入，仮名の適切な漢字への変換，翻字の誤りの修正，が含まれる．この内，可読性向上の効果が高く見込まれる文境界推定を本稿では扱う．

文境界推定は，主に音声認識の分野において，認識結果の文字列に句読点を挿入する目的で研究されてきた．古典的には，n-gram 言語モデル [1], [5],

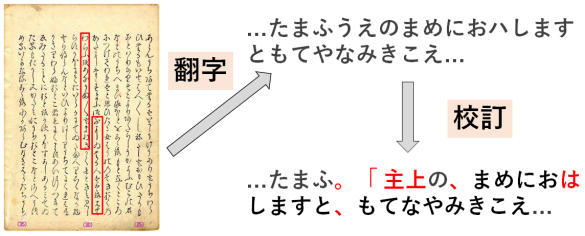

図 1 翻刻作業の工程

や，隠れマルコフモデル [3]，条件付き確率場（CRF）[7], [9], [14] が用いられた．深層学習の興隆以降は，LSTM（Long Short-Term Memory）[12], [16] や Bi-RNN（Bi-directional Recurrent Neural Network）[13] など文脈を考慮するモデルが提案されている．これらは，系列に句読点の有無をラベル付けするタスクとして文境界推定問題を解く．深層学習によるモデルの方が，入力の特徴空間をさらに非線形に写像した空間で識別境界を扱える分，性能が高くなる傾向がある．データが大量にある場合は，機械翻訳で用いられる RNN encoder-decoder モデルを用いて，句読点挿入前の文字列から挿入後の文字列に変換する手法も存在する [8], [15].

しかし，古典籍・古文書を対象にした文境界推定は筆者らの知る限り白井らの研究 [21] に限られる．白井らは，形態素に対して CRF を用いて文境界の有無のラベル付け行った．これは明治時代の雑誌が対象であるが，現代の文体との違いが大きい江戸時代以前の文献の方が，可読性向上の需要は高い．

そこで本稿は，平安時代の文献を対象に文境界推定を行う．深層学習の一種である Bi-LSTM（Bi-directional LSTM）[6] を用い，入力には形態素を利用する．精度の高い形態素解析には句点が事前に必要であることも考えられるが，本稿は自動形態素解析が人間の性能と同等であった場合を想定し，その上での文境界推定の性能を計測することに重点を置く．PR（Precision-Recall）曲線の AUC（Area Under Curve）で 0.894 を達成し，専門家による定性評価で高評価を得た．

2. 対象とするデータ

本稿は平安時代中期の中古日本語を対象とする．理由は次の三点である．まず，中古日本語の文章は，仮名が主であり，文境界検出による可読性の向上効果が高い．第二に，江戸時代まで中古日本語を引き継ぐ形で書き言葉が使われたため，後代の文語文へ拡張できる可能性が高い．第三に，コーパスが整備されており，計算機科学的な解析が行いやすい．

表 1 データの統計情報（文境界や記号は除く）

全文字数	全単語数	全文境界数	直後が文境界の単語の頻度
1,530,328	867,034	38,834	4.5%

一文中の単語数		
最頻値	最大値	最小値
16	334	2

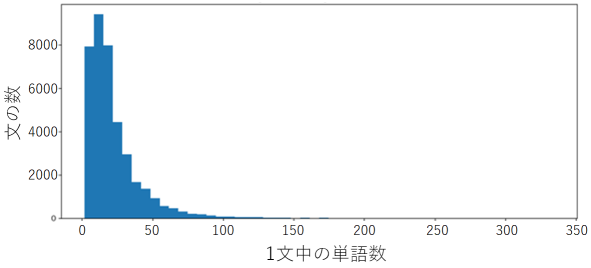

図 2 1 文中の単語数の分布

データセットには，日本語歴史コーパスの平安時代編 [17] を用いる．データは全て形態素解析済であり，自動形態素解析の後に手動で修正されている．コーパスの単語の単位は二種類あり，用例収集を目的とした短単位と，言語的特徴の解明を目的とした長単位がある．以下では短単位を「単語」とし，データ中の句点と空白（和歌の前後などに存在）を文境界とする．短単位の例としては，たとえば，次のようになる．むかし/男/初冠/し/て/奈良/の/京/春日/の/里/に/しる/よし/し/て/狩/に/いに/けり.

このコーパスに収録されているのは，古今和歌集，土佐日記，竹取物語，伊勢物語，落窪物語，大和物語，枕草子，源氏物語，紫式部日記，和泉式部日記，平中物語，堤中納言物語，更級日記，讃岐典侍日記，蜻蛉日記，大鏡，の 16 作品である．

データセットの統計情報（文境界や記号は除く）を表 1，1 文中の単語数の分布を図 2 に示す．表 1 に示すように，1 文中の単語数は，最大 334 単語，最小 2 単語で，最頻値は 16 単語である．図 2 に示すように，60 単語以下で 95.4%が網羅される．

3. 提案手法

文境界推定の手法として，Bi-LSTM [6] を用いた推定を提案する．使用できるコーパスのサイズが比較的小さいため，機械翻訳として解くよりも，ラベル付けのタスクとして解く方が性能が出る可能性が高いと予想される．Bi-LSTM の概要を図 3 に示す．Bi-LSTM は文頭から入力する LSTM と，文末から

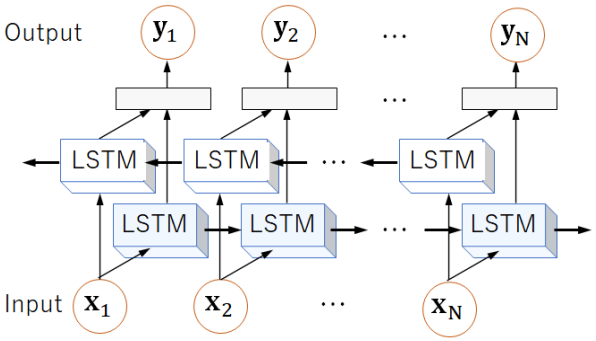

図 3 Bi-LSTM の概要

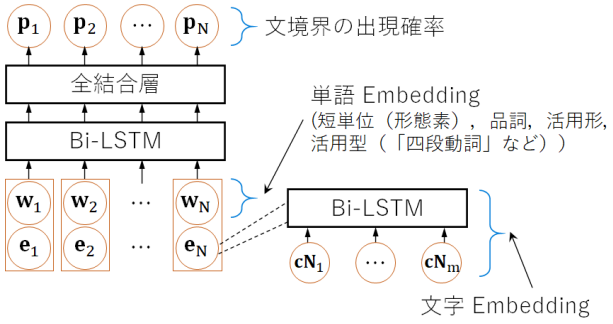

図 4 Bi-LSTM を用いた文境界推定モデルの概要

入力する LSTM の出力を結合し予測を行うモデルである．時系列データの特徴量を双方向から学習できるモデルで，前後の文脈を考慮できる利点がある．音声による文境界推定で高精度を誇り，中古日本語においても同様に機能することが期待できる．

Bi-LSTM に n 単語を入力し，各単語の直後に文境界が存在する確率を予測する．文境界推定の概要を図 4 に示す．入力には，単語の特徴ベクトル（形態素，品詞，活用形，活用型）に各単語の文字情報を連結して用いる．文字情報はそのまま単語の Embedding と連結するのではなく，Bi-LSTM を通したものを用いる．これを文境界推定用の Bi-LSTM に入力し，全結合層，softmax 関数を経て単語の直後が文境界である確率を出力する．表 1 にも示した通り，単語数と比較して文境界数は圧倒的に少ないため，境界部分に高い重みをかけて学習を行う．

ハイパーパラメータ調整は手動で行い，後述の評価手法において最も良かった値を採用した．例えば入力が 60 単語のとき，入力の次元数は 330（文字の情報が 100 次元，形態素が 200 次元，品詞・活用形・活用型を 10 次元ずつ），Bi-LSTM の中間出力は 100 次元，最終出力は文境界である確率の一次元である．境界部分にかける重みは 100 とした．最適化には SGD（Stochastic Gradient Descent）を使用し，学習率は 0.1 で，20 エポック反復し学習させた．フレームワークには PyTorch を用いた．

4. 実験

性能評価のため実験を行った．全単語列を入力単語数ずつ分割する．この分割したデータの中から，ランダムに取り出した 7 割を学習データに，残りの 3 割をテストデータに用いた．入力の単語数は，10，20，40，60 を試した．

文境界数は約 4 万で，音声認識での文境界推定（[15] では文境界数約 17 万）と比較して少なく，深層学習に十分なデータ量である保証がない．そこで比較手法として，連続する $2n$ 単語を入力し，その中心（n 単語目と $n+1$ 単語目の間）に文境界が存在するかどうかを推定させる決定木を Random forest [2] を用いて実装した．入力は Bi-LSTM と同様に単語の形態素・品詞・活用形・活用型を用い，文字情報は用いなかった．また，入力の単語数は 4，10，20 とした．これは，決定木の計算に必要なメモリ量の多さと，入力単語数の増加にともない精度が下がることによる．パラメータ調整をグリッドサーチにより行った．

評価関数には F1 値を用いた．これは，Precision（適合率）と Recall（再現率）の重みが等しい場合の調和平均である．Precision が高いほど検出した境界が正しく，Recall が高いほど境界を見落とさずに検出できることを意味する．Bi-LSTM は端に近い部分で精度が下がる傾向があるため，端から 3 分の 1 の結果は無視し，中心の 3 分の 1（入力が 60 単語の場合，中心の 20 単語）によってのみ評価した．

真陽性（True Positive, TP）が境界を正しく検出できた数に相当し，偽陽性（False Positive, FP）が境界でないのに境界と答えた数，偽陰性（False Negative, FN）が境界であるのに検出しなかった数である．Precision と Recall の計算式は次式の通りである．

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{1}$$

4.1 定量的評価

各手法の F1 値は，Random Forest を用いた比較手法において，入力が 4 単語のとき 0.83，10 単語のとき 0.81，20 単語のとき 0.77 であった．また，Bi-LSTM を用いた提案手法において，最高の F1 値は，入力が 10 単語のとき 0.82，20 単語のとき 0.81，40 単語のとき 0.82，60 単語のとき 0.83 であった．提案手法では，入力語数を増やすと精度が改善した一方で，比較手法では入力語数を増やすと精度が悪化した．

また，各手法で確信度を閾値として変化させた場

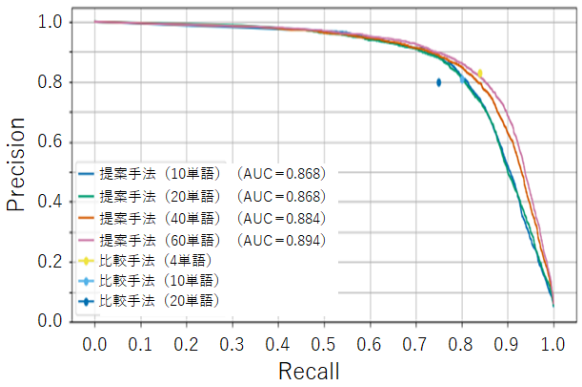

図 5　各手法の PR 曲線

表 2　失敗箇所の，直前の単語上位 4 種類

偽陽性		偽陰性	
ける	70 箇所	む	83 箇所
たまふ	33 箇所	や	62 箇所
けり	19 箇所	は	58 箇所
かな	15 箇所	ける	52 箇所

合の PR 曲線を図 5 に示す．Precision，Recall 共に高い方が良いため，カーブが右上に近いほど性能が高い．Random forest は決定木のため，閾値のようなパラメータがない．したがって，PR 曲線上では点で表現される．Bi-LSTM で 60 単語を入力した場合が最も性能が高く，PR 曲線の下側を積分した値である AUC は 0.894 と高い値となった．

提案手法の中でも，入力が 60 単語で最も F1 値が良かったとき，偽陽性（余計な文境界を入れた例）が 421 箇所（20.8%），偽陰性（文境界を見落とした例）が 1605 箇所（79.2%）と偽陰性の方が多かった．

4.2　エラー分析

失敗箇所の，直前の単語上位 4 種類を表 2 に，直前の単語の活用形上位 4 種類を，表 3 に示す．

偽陽性では，表 3 に示す通り，終止形の後に文が続くにも関わらず，直後に誤って文境界を入れてしまう例が多い．例を表 4 に示した．終止形の活用語の直後に助詞・助動詞が続くが，その前に文境界を入れた例があった．また，係り結び（連体形・已然形）があり，そこで文が終止すると判断してしまった例や，前に係助詞が存在しないのにも拘らず，連体形の直後に文境界を入れた例があった．

偽陰性では，表 3 に示したように，非活用語の直後に文境界を見落とす例が多い．例を表 5 に示した．間投助詞や詠嘆の終助詞の直後の文境界を見逃した例があり，これは文中の区切りではあるが文境界ではなく，読点を入れると良い例だと考えられる．また，係り結びを見落とした例があった．

表 3　失敗箇所の，直前の単語の活用形上位 4 種類

偽陽性		偽陰性	
終止形	251 箇所	非活用語	777 箇所
連体形	109 箇所	連体形	505 箇所
非活用語	44 箇所	終止形	129 箇所
已然形	13 箇所	已然形	110 箇所

表 4　提案手法の偽陽性の例 （予測：／，正解：／）

1) 終止形の活用語の直後に助詞・助動詞が続くが，その前に文境界を入れた例	… とばかり聞けばいみじうさきのごといさかふ ／ なり ／／ しばしありてはじめ…
2) 係り結びがあり，そこで文が終止すると判断してしまった例	… なされてしばしこそあれ ／ 一の宮の方に居させたまふ一品の宮后に立た…
3) 前に係助詞が存在しないにも拘らず，連体形の直後に文境界を入れた例	をねたむ女もありけり／／むかしものいひける ／ 女

表 5　提案手法の偽陰性の例 （予測：／，正解：／）

4) 係り結びを見落とした例	… 艶なう古めきたる直衣の裏表ひとしうこまやかなるいとなほなほしうつまづまぞ見えたる／ あさましと思す…
5) 間投助詞や詠嘆の終助詞の直後の文境界を見逃した例	… て荒らかにおどろおどろしく陀羅尼読むをいであな憎や／ 罪の深き身にやあらむ陀羅尼…

表 6　提案手法の成功例 （予測：／，正解：／）

偽陰性の失敗が多かった非活用語の直後で，文境界を正しく推定できた例	… 人急ぐとあるを今日は帰りて後にまゐりはべらむ ／／ そもそもかくてのみやはなど…
偽陽性・偽陰性共に失敗が多い係り結び周辺で，文境界を正しく推定できた例	… を長く折りて大きなる瓶にさしたるこそをかしけれ ／／ 桜の直衣に出袿してまらうど…

失敗が多く発生し，推定が困難であったと考えられるのは，非活用語の直後や係り結びの箇所である．しかし，これらの部分で成功している例も存在する．そのような例を表 6 に示す．

4.3　専門家からのフィードバック

中古日本語の作品の翻刻を頻繁に行う研究者 1 名に対し，実験結果について照会し，次のフィード

表 7 専門家へのヒアリングで使用された文境界推定の結果（予測：／，正解：／）

成功例
... て心ばへありてしつらひたり／／ 大和守のしわざなりけり／／ 人々もあざやかならぬ色の山吹掻練濃き衣青鈍などを着かへさせ薄色の裳青朽葉などをとかく紛らはして御台はまゐる／／ 女所にてしどけなくよろづのことならひたる宮の内にありさま心とどめてわづか...
は見えじとなむ思ひしかど人よりけにむつましうなりにたるこそとのたまはするをりをりはべり／／ くせぐせしくやさしだち恥ぢられたてまつる人にもそばめたてられではべらまし／／ 様ようすべて人はおいらかにすこし心おきてのどかにおちゐぬるをもととしてこそゆゑもよしもをかしく
御心本性の強くづしやかなるにはあらねど恥づかしげなる人の御気色のをりをりにまほならぬがいと恐ろしうわびしきなるべし／／ されど御硯などまかなひて責めきこゆればしぶしぶに書いたまふ／／ とりて忍びて宵の紛れにかしこに参りぬ／／ 大臣はかしこき
失敗例 （誤検出）
はちりかかるをやくもるといふらむ／ 家にありける梅の花の散りけるをよめるつらゆき暮ると明くと目かれぬものを梅の花いつの人まに移ろひぬらむ寛平御時后の宮の歌合の歌読人しらず梅が香を袖に移し
などかはむげにさしのぞかではあらむ／／ あやしからむ／女だにいみじう聞くめるものを／／ さればとてはじめつ方ばかりありきする人はなかりき／／ たまさかには壺装束などしてなまめき化粧じてこそはあめりしか／ それに物詣でなどをぞせし／／ 説経
おきてもゆかむ葦鶴の声振りいでてなきもとどめよ男返し難波江の潮満つまでになく鶴をまたいかなればおきてゆくらむあなそらごと／ つゆだに置かざめるものをとはいひけれどいかがありけむ／ その夜とどまりにけり／／ のちいかがなり
失敗例 （見落とし）
はべれば聞こえさせぬとのみあるにかやうの気色はさすがにすくよかなりとほほ笑みて恨みどころある心地したまふもうたてある心かな／ 色に出でたまひて後は太田の松のと思はせたることなくむつかしう聞こえたまふこと多かればいとどところせき心地して置き所なき
と思ひはかれどせちに否といふことなればえしひねば理なり／／ かぐや姫翁にいはくこの皮衣は火に焼かむに焼けずはこそまことならめと思ひて人のいふことにも負けめ／ 世になき物なればそれをまことと疑ひ
に女あさましくわびしうかなしうてただ泣きに泣かれていかに聞きたまひたるならむ／ いみじとはおろかなり／／ あこぎ惑ひ出でていかなることを聞こし召したるぞ／ さらにしあやまちせさせたまへることおはしまさざめるものをと申せばいでこの汐さきをかりてなさくじりそ／／

バックを得た．「翻字より校訂の方が習得に時間がかかる難しい作業であるが，それをこの精度で行えることに驚嘆する．失敗箇所も，ここには入れて欲しくない，といった失敗は見当たらず，言葉のまとまり単位でのミスであり，誤った箇所に入った文境界も邪魔にならない．連体形と終止形を取り違えるミスが多いが，これは読めばすぐに修正できる．今後は，句点だけでなく，読点や発話箇所の鉤括弧など，他の区切り文字も予測できるとなお読みやすくなるだろう．」この際に使用された，文境界推定の結果の一部を表 7 に示す．

5. まとめ・今後の課題

Bi-LSTM を用いて，平安時代の中古日本語の文境界を推定するモデルを構築し，PR 曲線の AUC で 0.894 と高精度な結果を得た．失敗している箇所は，終止形の直後に助詞や助動詞が続く場合や，係り結びの見落としなど，元々ある程度困難が予想される箇所であり，1 名の専門家からのフィードバックでは高評価を得た．今後は，係り結びの箇所での失敗などの頻発するミスを減らすようモデルの改善を行う．また，鍵括弧の挿入など，校訂における文境界推定以外の工程の自動化に着手する．

参考文献

[1] Doug Beeferman, Adam Berger, and John Lafferty. Cyberpunc: a lightweight punctuation annotation system for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 689—-692, 1998.

[2] Leo Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[3] Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. Punctuation annotation using statistical prosody models. In *Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40, 2001.

[4] Tarin Clanuwat, Alex Lamb, and Asanobu Kitamoto. Kuronet: Pre-modern japanese kuzushiji character recognition with deep learning. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[5] Agustin Gravano, Martin Jansche, and Michiel Bacchiani. Restoring punctuation and capitalization in transcribed speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4741—-4744, 2009.

[6] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[7] Madina Hasan, Rama Doddipatla, and Thomas Hain. Noise-matched training of crf based sentence end detection models. In *INTERSPEECH*, pages 349–353, 2015.

[8] Ondřej Klejch, Peter Bell, and Steve Renals. Sequence-tosequence models for punctuated transcription combining lexical and acoustic features. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[9] Wei Lu and Hwee Tou Ng. Better punctuation prediction with dynamic conditional random fields. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 177–186, 2010.

[10] Ayumu Nagai. On the improvement of recognizing single-line strings of japanese historical cursive. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 621–628, 2019.

[11] Hung Tuan Nguyen, Nam Tuan Ly, Kha Cong Nguyen, Cuong Tuan Nguyen, and Masaki Nakagawa. Attempts to recognize anomalously deformed kana in japanese historical documents. In *The 4th International Workshop on Historical Document Imaging and Processing*, 2017.

[12] Ottokar Tilk and Tanel Alumäe. Lstm for punctuation restoration in speech transcripts. In *INTERSPEECH*, pages 683—-687, 2015.

[13] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *INTERSPEECH*, pages 3047—-3051, 2016.

[14] Nicola Ueffing, Maximilian Bisani, and Paul Vozil. Improved models for automatic punctuation prediction for spoken and written text. In *INTERSPEECH*, pages 3097—-3101, 2013.

[15] Jiangyan Yi and Jianhua Tao. Self-attention based model for punctuation prediction using word and speech embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274, 2019.

[16] Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Ya Li. Distilling knowledge from an ensemble of models for punctuation prediction. In *INTERSPEECH*, page 2779–2783, 2017.

[17] 国立国語研究所（冨士池優美・須永哲矢・池上尚ほか）編. 『日本語歴史コーパス 平安時代編』（短単位データ 1.1 / 長単位データ 1.1, 中納言バージョン 2.2.0）. `https://pj.ninjal.ac.jp/corpus_center/chj/heian.html`, 2016.

[18] 国立歴史民俗博物館/東京大学地震研究所/京都大学古地震研究会. みんなで翻刻 - MINNA DE HONKOKU. `https://honkoku.org/`, 2020.

[19] 北本 朝展. データ駆動型人文学研究の発展と ai によるくずし字認識. **月刊** *J-LIS*, pages 36–39.

[20] 北本 朝展, カラーヌワット タリン, 宮崎 智, and 山本 和明. 文字データの分析—機械学習によるくずし字認識の可能性とそのインパクト—. **電子情報通信学会誌**, pages 563–568, 2019.

[21] 白井 良介, 松村 雪桜, 小木曽 智信, and 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. **情報処理学会論文誌**, pages 152–161, 2020.