契約書OCRの単語誤り訂正における漢字の偏旁冠脚を考慮した木編集距離の検討

阪本 浩太郎 †1 阿部川 明優 †2 佐竹 真樹 †2 岸川 至白 †2
阪本 エリーザ †4 石下 円香 †3 渋木 英潔 †3 森 辰則 †1

†1横浜国立大学 †2$GenialTechnology, Inc$ †3国立情報学研究所 †4無所属
E-mail: {sakamoto,mori}@forest.eis.ynu.ac.jp {aki,masaki.satake,itaru.kishikawa}@genialtech.io
sakamoto.e612@gmail.com {ishioroshi, shib}@nii.ac.jp

1 はじめに

会計業務の効率化を目的として，紙の契約書のスキャンデータから情報を自動抽出するシステムが求められている．我々は，契約書の構造により，会社名，住所，氏名などが書かれている範囲を推定し，推定された範囲ごとにOCRの結果からテキストを抽出するシステムを開発しているが，抽出されたテキストにOCRに起因する文字誤りが発生し，次のような漢字の偏旁冠脚置換や分離/融合による誤りが観察された．

- 「注文書」を「往文書」と誤認識
- 「会計」を「会言十」と誤認識

印字画像と類似した文字列であっても単語や文脈としては誤りとなるOCR誤りの訂正後処理は多く提案されているが，我々は漢字の偏旁冠脚誤りに注目し，氏名，会社名，住所，法律・会計用語等認識対象の種類ごとに予め用意した単語辞書のいずれかの見出し語に一致するようにOCR結果を自動訂正する手法を提案する．

OCR結果と見出し語の類似度を計測し最も類似する見出し語に訂正するためには，より見た目に沿った類似度を用いる必要があり，我々は，「注文書」と「往文書」のような漢字の偏旁冠脚置換誤りを訂正するための編集距離として漢字Damerau-Levenshtein距離（以降，漢字DL距離）を提案し，そこから算出される類似度を用いた訂正手法を提案した[1]．一般的な文字レベルの編集距離を用いるとOCR結果「往文書」から辞書の登録語「注文書」と「公文書」への編集距離が同一となってしまい，偏旁冠脚レベルで部分一致する「注文書」に訂正できないといった問題を漢字DL距離により解決した．しかし，漢字DL距離は「会計」と「会言十」のような偏旁冠脚が分離/融合してしまう誤りには適応できていなかった．また，漢字DL距離の計測に用いた偏旁冠脚データセット（漢字オンライン[1]の漢字構成）は12,300漢字に対応していたが，Unicodeには現在75,410漢字が登録されておりその多くに非対応であった．一方で，CHISE (Character Information Service Environment) プロジェクト[6]のIDS (Ideographic Description Sequence)[2]は，Unicodeの全漢字を偏旁冠脚に分解する規則であり，漢字をIDSの木構造（以降，IDS木）に変換することが可能である，従って，本研究では偏旁冠脚の置換と分離/融合を考慮可能なIDS木編集距離を提案し，それにより単語間類似度を計測しOCR文字誤り訂正を行う．

本研究では，IDS木編集距離による類似度と先行研究の漢字DL距離による類似度をそれぞれ用いた自動訂正結果を比較し報告する．

2 関連研究

OCR誤り訂正として，文脈的な大域情報を考慮した訂正手法[2][3]や予め用意した単語辞書のいずれかの見出し語に一致するように訂正する手法[4][5]が多く提案されている．我々の研究では，大局情報を用いることができないため単語辞書ベースの訂正を行い，OCR結果に含まれる漢字を偏旁冠脚にまで分解して誤り訂正を行う点が他の手法と異なる．

3 本研究の対象

想定する全体の処理の流れを図1に示す．OCR結果と，氏名，会社名，住所，用語といったデータ型を入力し，OCR文字誤り検出を行う．誤りが検出されない場合はOCR結果をそのまま出力する．誤りが検出された場合は，OCR文字誤り訂正を行う．OCR文字誤り訂正では辞書からOCR結果と最も類似する登録語を取り出し，それを訂正結果として出力する．OCR文字誤り訂正はデータ型によって処理の流れが異なる．

本稿では，OCRの誤りが検出された場合のOCR結果の訂正部分のみに着目する．

[1] https://kanji.jitenon.jp/

[2] http://git.chise.org/gitweb/?p=chise/ids.git

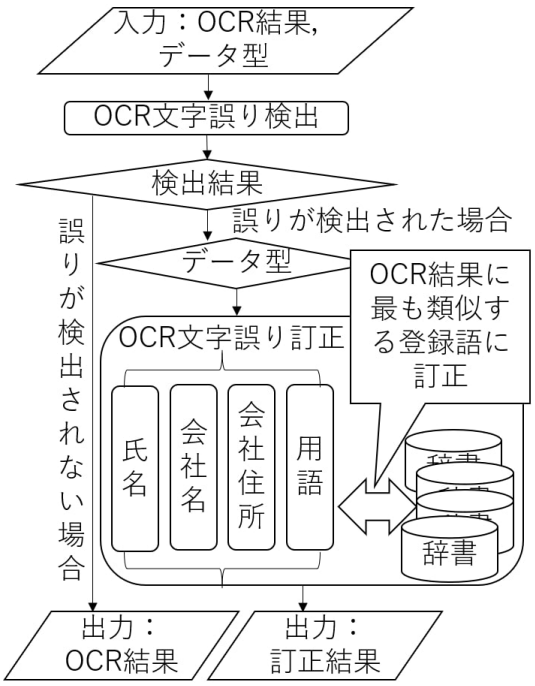

図 1: 全体の処理の流れ

4 辞書データ

本章では，作成した辞書データについて説明する．

4.1 人名辞書

人名辞書を NEologd[3]の辞書から作成した．重複を削除し，アルファベット含むものとカタカナのみのものも削除した．登録した苗字は 74,744 件，名前は 82,704 件である．なお，これらの内，ひらがなのみで構成された苗字は 0 件，名前は 408 件である．

4.2 法人名・法人住所辞書

法人名とその住所を登録した法人名・法人住所辞書を法人登記データ（令和元年 8 月 30 日更新版）[4]から作成した．登録した法人数は 4,783,578 件である．法人登記データから得られた法人情報は，法人住所を，都道府県，市区町村，番地・建物の 3 つに分解した上で，法人名から番地・建物，番地・建物から市区町村，市区町村から都道府県への参照を残す形で，それぞれ辞書に登録した．

4.3 法律・会計用語辞書

「有斐閣法律用語辞典第 4 版」[5]，政府機関「日本法令外国語訳推進会議」による「法令用語日英標準対訳辞書」[6]，経営者と士業者のマッチングサイト「士業ねっと！」による「会計用語キーワード辞典」[7]のそれぞれの辞書の見出し語から合計 16,009 語を法律・会計用語辞書に登録した．

5 予備調査（OCR の選定）

本研究で使用する OCR を選定するため，OCR の性能を調査した．人名辞書，法人名・法人住所辞書，法律・会計用語辞書から氏名，法人名，法人住所，用語をそれぞれ 1,100 件ずつをランダムに選択し，10.5pt で紙に印刷し，原則 300dpi でスキャンし，次の 5 つの OCR の結果を比較した．Genial Technology, Inc.（以降，GT 社）で現在開発中の OCR，LSTM に基づく OCR である Tesseract [8]（4.1.0 版）[9]，Adobe Acrobat Pro DC (Continuous Release のバージョン 19.012.20040) に搭載されている OCR，富士ゼロックス社のプリンタ複合機 ApeosPort-VI C5571 に搭載されている OCR，Google Cloud Vision API[10]（2019 年 12 月 4 日時点）を用いた．

表 1 は各 OCR の出力結果が元の単語と一致するかを表す正答率であるが，GT 社の OCR の性能が最も良いことが読み取れる．

さらに，辞書登録されている全見出し語から訓練データを作成し，Tessearct でファインチューニングを行い（イテレーション 10,000 回），人名辞書と法律・会計用語辞書に特化した OCR をそれぞれ作成した．それらの OCR をスキャンデータ各 20,000 単語に対し使用し，上述と同様の方法で調べた結果が表 2 である．辞書特化 OCR と比較しても GT 社の OCR の方が性能が良いことがわかった．

以上の結果により，本研究では GT 社の OCR を用いて OCR 誤りを自動訂正する手法を提案する．

表 1: OCR の正答率 (/1,100 単語)

OCR	氏名	法人名	法人住所	用語
GT	0.885 974	**0.976** **1,074**	**0.984** **1,082**	**0.992** **1,091**
A	**0.945** **1,040**	0.860 946	0.905 996	0.940 1,034
B	0.901 991	0.945 1,039	0.975 1,073	0.991 1,090
C	0.901 991	0.885 974	0.920 1012	0.963 1059
D	0.596 656	0.821 903	0.783 861	0.927 1,020

[3]https://github.com/neologd/mecab-ipadic-neologd
[4]http://www.houjin-bangou.nta.go.jp/download/zenken/
[5]http://www.yuhikaku.co.jp/dictionary/detail/9784641000285
[6]http://www.japaneselawtranslation.go.jp/dict/download/
[7]https://kaikei-yougo.sigyo.net/
[8]https://github.com/tesseract-ocr/tesseract/
[9]モデルは https://github.com/tesseract-ocr/tessdata_best/ の jpn.traineddata を使用
[10]https://cloud.google.com/vision/

表 2: 辞書特化 OCR との正答率比較 (/20,000 単語)

OCR	氏名	用語
GT	**0.885**	**0.995**
	17,690	**19,895**
辞書特化	0.642	0.942
	12,842	18,834

6 IDS 木

IDS とは漢字の漢字構成への変換をポーランド記法で記述する形式であり，演算子は Unicode の漢字構成記述文字で表現される．図 2 に IDS の 3 例を表示する．&CDP-89BB; は実体参照形式で書かれた 1 文字であり，CDP は台湾中央研究院 CDP 外字を指し，89BB はそのコードポイントである．図 3 のように，IDS による変換を漢字構成の要素に対して変換が恒等写像（図中の赤色の変換）となるまで繰り返し適用することで偏旁冠脚を葉とする変換木が作成され，これを IDS 木と呼ぶ．

7 単語間の類似度による自動訂正

本章では，図 4 の OCR 結果の自動訂正手法について説明する．まず，データ型と誤りを含む OCR 結果を入力し，次にデータ型に対応する辞書の見出し語リストを取得する．次に OCR 結果と見出し語間の編集距離を測定し，そして編集距離から類似度へ変換し，最後に類似度に基づき OCR 結果を訂正する．編集距離の測定方法を 7.1 節と 7.2 節，編集距離から類似度への変換を 7.3 節，類似度による単語の訂正を 7.4 節で説明する．

7.1 IDS 木編集距離

Zhang-Shasha 木編集距離は，1 つの（1 根順序付き）木を別の木に変形するのに必要な操作の最小回数である．操作は，挿入，削除，交換の 3 種類がある．IDS 木編集距離は，2 つの単語に対して，単語に含まれる漢字を IDS 木に変換して作成される IDS 木の Zhang-Shasha 木編集距離である．図 5 は「ヘッジ会計」と「ヘッジ会言十」の IDS 木であり，赤い囲いのノード 1 つの削除/挿入操作 1 回が編集距離となる．漢字を IDS 木の葉の列ではなく木構造で扱う理由は，例えば「森」と「木林」という苗字の葉の列はどちらも「木木

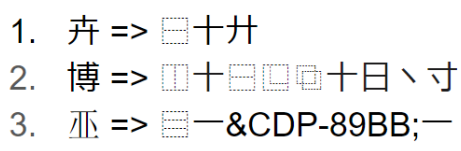

図 2: IDS 形式による漢字から漢字構成への変換例

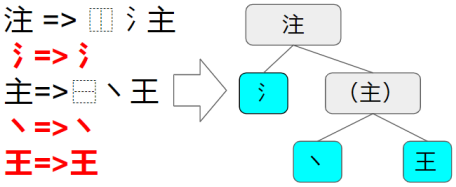

図 3: IDS 木の作成

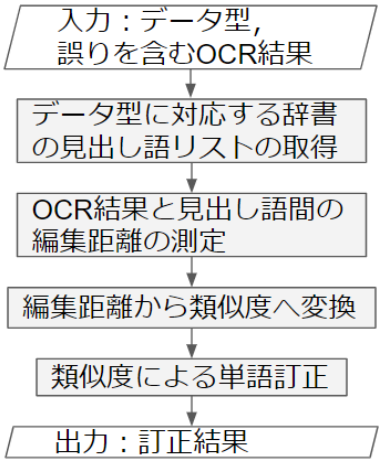

図 4: OCR 結果の自動訂正の処理の流れ

木」となって一致してしまうが，木構造で扱うことで一致させずに距離が測れるからである．

7.2 漢字 DL 距離

Damerau-Levenshtein 距離は，1 つの文字列を別の文字列に変形するのに必要な操作の最小回数である．操作は文字の挿入（insertion），削除（deletion），置換（substitution），隣接文字の交換（transposition）の 4 種類がある．漢字 DL 距離は，OCR による漢字誤り訂正を目的として，漢字置換の際に，偏旁冠脚の部分一致や画数といった内部情報の類似性の分だけ漢字置換コストを和らげ距離が縮まるよう Damerau-Levenshtein 距離を拡張したものである．

7.3 編集距離から類似度への変換

IDS 木編集距離や漢字 DL 距離から類似度への変換は次式で行う．

$$sim(s1, s2) = 1 - \frac{distance(s1, s2)}{max(length(s1), length(s2))}$$

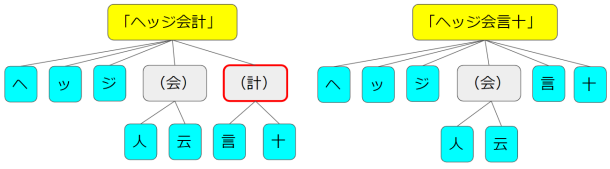

図 5: IDS 木編集距離

表 3: 実験結果の正答率（/20,000 単語）

訂正手法	氏名	法人名	法人住所	用語
漢字 DL 距離	0.889 17,783	0.947 18,938	0.968 19,359	0.998 19,959
(none)	0.885 17,690	0.934 18,682	0.968 19,359	0.995 19,895

7.4 類似度による単語の訂正

法人名と用語の訂正は OCR 結果と最も類似する会社名を辞書から選択して行う．氏名と法人住所の訂正は，OCR 結果の先頭から末尾まで順次，辞書から単語を選択し訂正を行う．

8 実験

提案した誤り訂正手法の性能を比較評価するため，OCR 結果がすべて誤りであると検出された前提で，誤り訂正を行った結果に対し，正答率で評価する．正答率は，訂正結果と正解情報の間の文字列の表層完全一致による一致率で計算する．IDS 木編集距離と漢字 DL 距離に基づく類似度をそれぞれ用いて誤り訂正する．使用する辞書については，4 章で述べた辞書のうち，人名辞書と法律・会計用語辞書はそのまま用いたが，法人名・法人住所辞書については登録されている件数が多く処理に時間が掛かりすぎるため，代わりに OCR データセットの正解情報のリストを簡易辞書として用いて実験を行った．実験に用いた OCR 結果は，人名辞書，法人名・法人住所辞書，法律・会計用語辞書から氏名，法人名，法人住所，用語をそれぞれ 20,000 件ずつをランダムに選択し，10.5pt で紙に印刷し 300dpi でスキャンし OCR にかけた．フォントは MS 明朝，MS P 明朝，メイリオ，游ゴシックを等頻度で使用した．

何も行わず OCR 結果をそのまま出力した場合の正答率とも比較する．

9 結果と考察

漢字 DL 距離についての実験の結果を表 3 に示す．IDS 木編集距離については，計算時間がかかりすぎたため，代わりに OCR 誤りの末尾 10 件に対してのみ訂正を行った．その結果，氏名は 1 件，法人名は 10 件，法人住所は 7 件，用語は 9 件，正しく訂正できた．IDS 木編集距離によって，OCR 結果「素乱」を「紊乱」に訂正する偏旁冠脚誤り訂正に成功していることを確認した．

10 まとめと今後の課題

会計業務の効率化を目的として，紙の契約書のスキャンデータから情報を自動抽出するシステムの改良に向け，OCR により抽出されたテキストの OCR 文字誤りを自動訂正する手法を提案した．偏旁冠脚に注目した IDS 木編集距離を提案し調査を行ったが，実行時間を多く要したため本稿では簡易な調査の結果を報告した．今後は IDS 木編集距離の調査を最後まで実行して比較分析を行う予定である．また，実行時間を短くする方法についても調査する．

参考文献

[1] 阪本 浩太郎, 阿部川 明優, 岸川 至白, 阪本 エリーザ, 石下 円香, 渋木 英潔, 森辰則. 契約書の OCR 漢字誤り訂正における偏旁冠脚を考慮した編集距離の検討. 自然言語処理研究会報告 2019-NL-242, 情報処理学会,2019.

[2] Masaaki Nagata.: *Japanese ocr error correction using character shape similarity and statistical language model* Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, pp. 922–928, 1998.

[3] 増田 勝也. 大域的情報を用いた OCR 文字誤り訂正. 言語処理学会 第 21 回年次大会, pp. 127–130, 2015.

[4] 寺崎 正則，清野 和司，山城 さつき．自由記載姓名文字列に対する知識処理．全国大会講演論文集第 41 回 (人工知能及び認知科学), 208-209, 1990-09-04

[5] 丸川 勝美, 古賀 昌史, 嶋 好博, 藤沢 浩道．手書き漢字住所認識のためのエラー修正アルゴリズム. 情報処理学会論文誌 35(6), 1101-1110, 1994-06-15.

[6] Tomohiko Morioka, *Multiple-policy Character Annotation based on CHISE* In Journal of the Japanese Association for Digital Humanities, 2015 Volume 1 Issue 1 Pages 86-106

[7] Kaizhong Zhang and Dennis Shasha. *Simple fast algorithms for the editing distance between trees and related problems* SIAM Journal of Computing, 18:1245–1262, 1989.