日本語情報抽出タスクのための LayoutLM モデルの評価

西脇 一尊 [1]　大沼 俊輔 [1]　門脇 一真 [1]
[1] 株式会社日本総合研究所
{nishiwaki.kazutaka,onuma.shunsuke,kadowaki.kazuma}@jri.co.jp

概要

本研究では，文書のレイアウト情報を活用して学習・推論を行う LayoutLM を日本語コーパスを用いて学習を行い，情報抽出タスクにおける性能の検証を行った．実験では Wikipedia 記事を用いて LayoutLM の事前学習を行い，2 種類の情報抽出タスクのための Fine-tuning を行った．ベースラインの BERT との性能比較を行った結果，一方のタスクにおいてレイアウト情報が性能向上に寄与することが確認できた．本稿では一連の学習と評価タスクにおける LayoutLM と BERT の性能について報告を行う．

1　はじめに

テキストから意味内容を構造化された形式に変換する情報抽出技術は，日々大量に作成・更新されるテキストに含まれる情報を利活用する上で重要である．電子媒体や紙媒体といった媒体を問わず，文書には人間が情報を理解しやすくするためのレイアウトが施されている場合が多い．しかし，多くの情報抽出に関する研究ではテキスト内容から学習した言語モデルを用いており，特に日本語におけるレイアウト情報を活用した情報抽出の研究は，英語を対象とした研究と比べ活発には行われていない．

本研究では，LayoutLM [1] を日本語コーパスを用いて事前学習を行い，BERT [2] との情報抽出タスクにおける性能を比較することで，日本語の情報抽出タスクにおける LayoutLM の性能を評価する．

2　関連研究

本研究で取り扱うレイアウト情報とは，文書に含まれる単語のその文書中での位置を指す．以下では，レイアウト情報を考慮できる言語モデルである LayoutLM の概要と，情報抽出に関する取り組みの例として，Wikipedia 記事からの情報抽出・構造化を目的とするプロジェクトについて述べる．

2.1　LayoutLM

LayoutLM [1] は，レイアウト情報を考慮できる言語モデルの 1 つとして知られている．これは BERT [2] のアーキテクチャにトークンの文書中における位置を表現する 2-D Position Embedding を追加することで，文書のレイアウト情報を考慮できるようにした言語モデルである．LayoutLM では，

1. マスクされたトークンを文脈とレイアウト情報から推論を行う Masked Visual-language Model (MVLM)
2. 文書画像をマルチラベル分類[1)]することで文書レベルの特徴を学習する Multi-label Document Classification (MDC)

の 2 つの手法を用いて事前学習を行っているが，MDC は必須ではないとされている．

レイアウト情報を学習した LayoutLM は，スキャンされた文書やレシート等を元にしたデータセットにおいて Semantic labeling や Slot filling の性能が向上することが報告されており，日本語においても帳票や請求書での情報抽出の精度が既存手法よりも向上することが報告されている[2)] [3)]．しかし，日本語においては一般利用可能な LayoutLM の事前学習済みモデルが公開されておらず，観測できる範囲では LayoutLM を活用した取り組み事例は限られている．

2.2　森羅プロジェクト

森羅プロジェクト [3] は，Wikipedia に記載されている知識を計算機が扱えるよう構造化することを目的としているプロジェクトである．7 回目の評価タスク開催となる森羅 2022[4)] では，

1. Wikipedia 記事を拡張固有表現 [4] の階層定義に

1） ラベルの一例：メール，ニュース記事，請求書，論文
2） https://cinnamon.ai/ideas/2021/01/18/20210118_research_006/
3） https://tech.layerx.co.jp/entry/2022/11/24/130000
4） https://2022.shinra-project.info

基づき定義されたカテゴリに分類するタスク
2. カテゴリ毎に定義されている属性値を抽出するタスク
3. 属性値テキストと意味的に一致する Wikipedia 記事との紐付けをするタスク

の 3 タスクが実施されている．本研究では 2 の属性値抽出タスクを LayoutLM の性能評価に用いる．

3 評価タスク設定

日本語での LayoutLM の性能を評価するにあたり，以下の 2 種類の評価タスクを設定した．以下ではこれらのタスクについて述べる．

3.1 森羅 2022 - 属性値抽出タスク

属性値抽出タスクは，カテゴリ別に分類された Wikipedia 記事から各カテゴリ毎に定義された属性に基づいて，その特徴を表す属性値を抽出するタスクである．例えば，「スターバックス（カテゴリ：企業名）」の記事から，「創業国」の属性値として“アメリカ合衆国”というテキストを抽出する．このタスクは固有表現抽出と類似しているが，異なる点として抽出対象の属性がカテゴリや文脈に応じて変化することが挙げられる．例として，「ホセ・コントレラス（カテゴリ：人名）」の記事では，“アメリカ合衆国”のテキストは文脈に応じて「国籍」または「居住地」の属性値とされる．また，属性はカテゴリを跨いで重複する場合もあり，例えば「地位職業」という属性は「人名」と「キャラクター名」の 2 種類のカテゴリにおいて設定されている．

森羅 2022 の属性値抽出タスクは，属性値抽出の対象となる 178 カテゴリに該当する Wikipedia 記事から属性値を抽出するタスクとなっている．学習用データでは重複を除くと計 1,700 種類の属性が使用されている．アノテーションされた記事の件数は 1 カテゴリあたり平均 111 件である．

3.2 契約書からの情報抽出

Web ページ以外の文書を対象とする情報抽出性能について評価するために，企業間の契約書中に記載されている契約締結日，契約開始・終了日，契約締結社名等の 7 種類の情報を抽出するタスクを設定した．

このタスクで用いるデータセットは，契約書データを取り扱う企業から提供された機密情報を含まないテンプレートをコーパスとしている．このコーパスに対して社名・日付・契約内容といった記載が必要な部分に対してアノテーションを行い，ランダムに選択した実在する企業名やダミーの日付・金額表現等を挿入することで独自に作成されたものである．このデータセット（以後，ダミー契約書）の特徴として，「契約金額」を指し示す表現として「金額」「委託料」「請求金額」などの表記揺れや，「業務委託契約」「請負契約」「秘密保持契約」等複数の契約書の種別が存在することが挙げられる．

4 実験手順

4.1 事前学習データセット

森羅 2022 で配布された 2019 年版 Wikipedia の全記事データ（約 110 万記事分）を用いた．トークンと座標情報の系列を取得するために，まず各記事データの HTML をブラウザで描画[5]し，その画面に表示されたテキストを別途トークナイズしてから，各トークンのブラウザ上での描画位置（座標情報）を取得した[6]．また，LayoutLM ではレイアウト情報を 0 から 1,000 の範囲に正規化する必要があるため，併せて body 要素全体の座標情報も取得している．

4.2 Fine-tuning データセット

属性値抽出タスク　森羅 2022 で配布されている 2019 年版 Wikipedia 記事の一部と，対応する属性値のアノテーションを用いた．

事前学習データセットの構築時と同様の手順で，各記事からトークンと座標情報の系列を取得した．本実験では系列ラベリング問題として属性値抽出タスクを解くため，属性値のアノテーションをオフセット形式から IOB2 形式に変換し系列に組み込んだ．この時，オフセット区切りとトークンの区切りが合致しない場合，オフセット区切り位置を含む（より長い範囲が選択される）ようラベルを作成している．また，同じテキストに複数のアノテーションがされている場合，配布されたアノテーションファイルを先頭から読み込んだ際に最初に参照する属性値を採用した．

アノテーションデータは 9:1 の割合で分割し，それぞれを学習・検証データとして用いた．テストデータとして，森羅 2022 評価タスクのリーダー

5） 描画環境：Google Chrome 106.0.5249.119，Ubuntu 20.04.5 LTS，画面サイズ 1,280*854
6） 事前に各トークンを nowrap 値を指定した span 要素で囲み，各要素の BoundingClientRect を座標情報として用いた

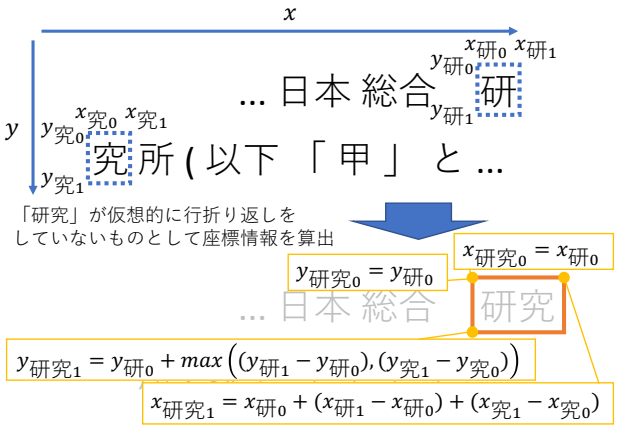


図 1: 「研究」が 1 つのトークンであり，「研」と「究」の間に行折り返しがされた際の「研究」トークンの座標情報の取得イメージ

ボード提出用として指定されている 2021 年版の Wikipedia 記事に対して，同様の手順でトークンと座標情報の系列を抽出したものを用いた．このテストデータの正解データは非公開であるため，テストデータでの性能評価はリーダーボード上の公開スコアをもとに行った．

各ページの系列長が LayoutLM の上限入力長である 512 トークンを超える場合，一定の重複範囲（本実験では 128 トークンと設定）を持たせたスライディングウィンドウによりデータを分割した．

契約書からの情報抽出タスク　3.2 節で作成したダミー契約書の PDF は，PyMuPDF[7] を用いて，

1. テキスト全文
2. 文字単位の座標情報と正解ラベル[8]
3. 契約書 1 ページ毎のサイズ

の 3 点を抽出した．1 のテキスト全文はトークナイズし，トークンに含まれる各文字の座標情報からトークンの座標情報を求めた．トークン中に行の折り返しが生じた場合には，折り返し後に続く文字の高さと幅から折り返しがなかった時の座標情報を算出した．このときの座標情報の取得イメージを図 1 に示す．

本実験では系列ラベリング問題として契約書からの情報抽出タスクを解くため，トークンを構成する文字に正解ラベルが付与されている場合，対応するラベルに B タグ・I タグを設定することで IOB2 形式のラベルを作成した．また，トークンの座標情報は 3 の契約書ページサイズを用いて 0 から 1,000 の範囲に正規化をした．

データセットは 8:1:1 の割合で分割し，それぞれを学習・検証・テストデータとして用いた．

属性値抽出タスクのデータセットと同様に，系列長が 512 トークンを超える場合はスライディングウィンドウによりデータを分割した．

4.3　LayoutLM の事前学習

BERT-base に相当するパラメータサイズの LayoutLM の事前学習を行った．BERT と共通するパラメータは東北大 乾研究室が公開している日本語版 BERT[9] の重みで初期化し，2-D Position Embedding はランダムな値で初期化した．

評価タスクにて用いる Wikipedia やダミー契約書は，文書毎に大きなレイアウトの差異がなく情報抽出に有用な特徴を学習することは難しいと考えたため，本実験では MVLM のみ実施し MDC は実施しなかった．

ハイパーパラメータは学習ステップ数 100,000，学習率 5e-5，バッチサイズ 256 とした．なお，学習率は 1,000 ステップまで Warmup を行い，そこから線形に減衰させた．

学習は 8 基の NVIDIA A100 SXM4 を用いて行っており，後述の Fine-tuning や推論にも同環境を用いた．

4.4　評価のための Fine-tuning

4.3 節で学習した LayoutLM とベースラインの BERT の Fine-tuning は，Hugging Face [5] が公開しているスクリプト[10] をベースに実装した．いずれのモデルも，ハイパーパラメータはグリッドサーチ[11] を行い，検証データにおいて最良の micro-F1 を達成したモデルを評価に採用した．

属性値抽出タスクにおいて，LayoutLM と BERT の性能差が Wikipedia 記事を用いた学習ステップ数の差によるものではないことを確認するため，事前学習済み BERT に加え，2019 年版 Wikipedia 記事を用いて 100,000 ステップの Masked Language Model [2] による追加学習を行った BERT+100K モデルを用意

7）https://pymupdf.readthedocs.io/en/latest/
8）3.2 節でダミーデータを挿入する際に情報種類毎に異なる色を指定しておき，PyMuPDF で取得した色情報をもとに文字毎の正解ラベルを判定した．なお，色情報はモデルの学習や推論に利用しない
9）https://huggingface.co/cl-tohoku/bert-base-japanese-v2
10）https://github.com/huggingface/transformers/blob/v4.20.1/examples/pytorch/token-classification/run_ner.py
11）学習率:3e-5,4e-5,5e-5,6e-5, バッチサイズ:8,16,32, エポック数:5, 10, 15, 20

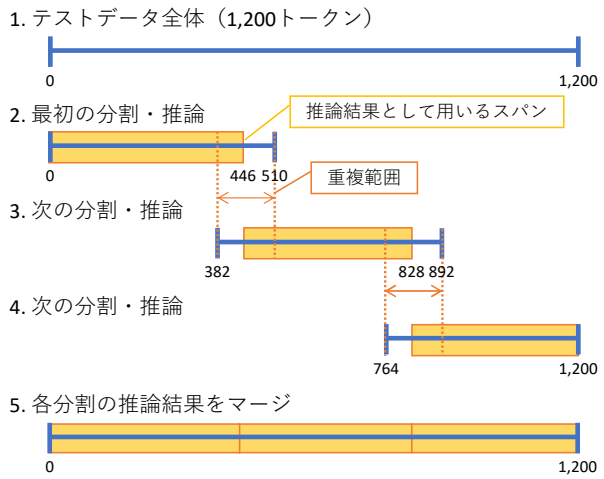


図 2: 1,200 トークン長のテストデータをスライディングウィンドウで分割し，推論結果をマージする処理のイメージ．ウィンドウ幅はモデルに入力する際の [CLS] と [SEP] 分を差し引いた値である

し，3 つのモデルそれぞれで Fine-tuning を行い性能を比較した．

4.5 テストデータでの推論

スライディングウィンドウによりテストデータが分割された場合，ウィンドウに含まれる重複範囲は 2 回推論が行われる．重複範囲の前半は前側ウィンドウの，後半は後ろ側のウィンドウの推論結果を採用し，最終的に各ウィンドウでの推論結果をマージして推論結果としている（図 2 参照）．

また，属性値抽出タスクにおいて，3 モデルの推論結果いずれにおいても以下の後処理を行った．

- 推論結果 top-5 のうち，カテゴリに存在しない属性を無視した上で logit が最大の推論結果，または O タグを出力する
- 同じ属性の I タグ同士の間に 5 トークン未満の連続した O タグが含まれている場合，O タグを I タグに変換する[12)]
- O タグ，もしくは異なる属性の B または I タグに続く I タグを B タグに変換する

5 実験結果

属性値抽出タスク 属性値抽出タスクにおける各モデルの検証データによる実験結果，およびテストデータによる推論結果の公開スコアを表 1 に示す．

契約書からの情報抽出タスク 契約書からの情報抽出タスクにおける各モデルの検証・テストデータによる実験結果を表 2 に示す．

12） 属性値抽出タスクにおいて，属性値中の記号や助詞等が O タグとされる傾向があったため

表 1: 属性値抽出タスクにおける実験結果．太字は最良のスコアを表す

	検証データ	テストデータ	
手法	micro-F1	macro-F1	micro-F1
LayoutLM	**64.37**	**55.15**	**63.00**
BERT	63.03	54.75	60.14
BERT+100K	63.27	54.82	61.28

表 2: 契約書からの情報抽出タスクにおける実験結果．太字は最良のスコアを表す

	検証データ	テストデータ
手法	micro-F1	micro-F1
LayoutLM	97.88	98.28
BERT	**97.94**	**98.69**

6 考察

属性値抽出タスクにおいて LayoutLM はベースラインモデルの性能を上回っているが，これは LayoutLM がレイアウト情報からトークン列が同じ行や箇条に記述されていることを手がかりに推論することができた可能性がある．レイアウト情報が抽出に影響を及ぼしたと考えられる例を付録 A に記載する．

契約書からの情報抽出タスクにおいて LayoutLM はベースラインモデルの性能をわずかに下回ったが，これは LayoutLM の事前学習を Wikipedia コーパスを用いて行ったためモデルが Wikipedia 記事に過剰に適合してしまい，ダミー契約書での性能が BERT と比べて劣った可能性がある．

7 おわりに

本研究では，日本語 Wikipedia 記事を用いて LayoutLM の事前学習を行い，2 種類の日本語情報抽出タスクにおける LayoutLM の性能を調査した．LayoutLM は，契約書からの情報抽出では期待された性能向上が見られなかったが，Wikipedia 記事からの属性値抽出タスクの性能が向上したことが示され，日本語文書における情報抽出のアプローチの 1 つとして有望であることがわかった．

今後の展望として，LayoutLMv2 [6] 等の画像情報も考慮して学習するモデルの利用について検討する．また，事前学習のデータセットを契約書等の文書に変更した際の性能評価や，有価証券報告書等のレイアウト情報が情報抽出に寄与する可能性のある文書においてのモデルの性能を調査する．

謝辞

本研究は産総研の AI 橋渡しクラウド（ABCI）を利用したものです.

参考文献

[1] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '20, pp. 1192—1200, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[3] Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. SHINRA: Structuring wikipedia by collaborative contribution. In **AKBC**, 2019.

[4] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**, 2002.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

[6] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 2579–2591, 2021.

A　付録

Wikipedia からの属性値抽出タスクにおいて，レイアウト情報が抽出においてポジティブ・ネガティブに影響を及ぼしたと考えられる例を示す．なお，付録中の図に含まれるスクリーンショットの一部は Google Chrome 上で幅 1,280px で描画したものを用いている．

A.1　レイアウト情報がポジティブに影響を及ぼしたと考えられる例

「Adobe Director」『フリー百科事典　ウィキペディア日本語版』．2017 年 3 月 18 日 (土) 11:14 UTC の一部スクリーンショットを図 3a と，ハイライト箇所における系列ラベリングの結果を図 3b に示す．系列ラベリング結果中のハイライト箇所の通り，LayoutLM のみ正しく“Macromedia Director 7 Shockwave Studio”を単一の属性値として抽出できた．これは，レイアウト情報から Shockwave Studio のテキストが前方のテキストと同じ行であることを理解していることから一続きの属性値として抽出できた可能性がある．

バージョン履歴[編集]

- 1985年 **MacroMind**VideoWorks
- 1987年 MacroMind VideoWorks II
- 1988年 MacroMind Director 1.0
- 1988年 MacroMind Director 2.0
- 1989年 MacroMind Director 3.0
- 1993年 **Macromedia** Director（バージョン3.1.3）
- 1993年 Macromedia Director 4
- 1996年 Macromedia Director 5
- 1997年 Macromedia Director 6
- 1998年 Macromedia Director 7
- Macromedia Director 7 Shockwave Studio
- Macromedia Director 7 Lite
- 2000年 Macromedia Director 8

(a)「Adobe Director」のスクリーンショット一部（ハイライトによる強調は著者によるもの）

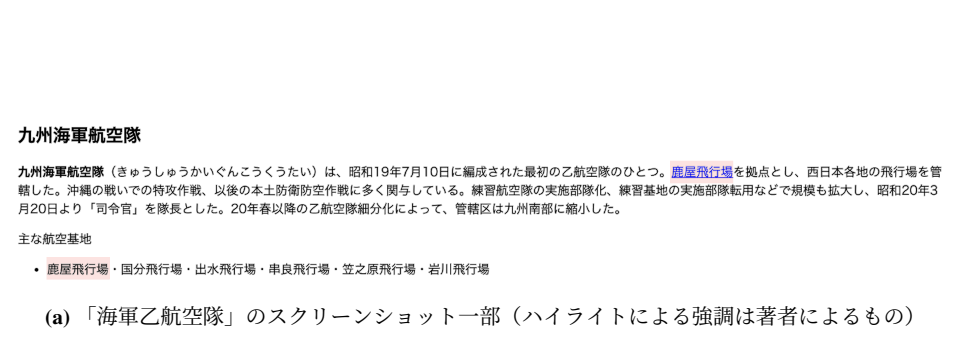
トークン	真のラベル	予測ラベル		
		LayoutLM	BERT	BERT+100K
Mac	B-バリエーション	B-バリエーション	B-バリエーション	B-バリエーション
##rom	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##edia	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Direct	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##or	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
7	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Sh	I-バリエーション	I-バリエーション	B-バリエーション	B-バリエーション
##ock	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##w	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##ave	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Studio	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Mac	B-バリエーション	B-バリエーション	B-バリエーション	B-バリエーション
##rom	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##edia	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Direct	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##or	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
7	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
Li	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション
##te	I-バリエーション	I-バリエーション	I-バリエーション	I-バリエーション

(b) ハイライト箇所に対する系列ラベリング結果

図 3: レイアウト情報が抽出性能にポジティブに影響を及ぼしたと考えられる例

A.2　レイアウト情報がネガティブに影響を及ぼしたと考えられる例

「海軍乙航空隊」『フリー百科事典　ウィキペディア日本語版』．2018 年 5 月 9 日 (水) 07:34 UTC の一部スクリーンショットを図 4a と，ハイライト箇所における系列ラベリングの結果を図 4b に示す．系列ラベリング結果中のハイライト箇所の通り，LayoutLM のみ文章中に現れる“鹿屋飛行場”を属性値として抽出できなかった．一方で，同ページ内の左寄せされた“鹿屋飛行場”は正しく抽出できている．これは，学習データに含まれる「基地」の属性値が左寄せの箇条書きで書かれることが多く，レイアウト情報からページの右側に寄ったテキストを属性値として認識できていない可能性がある．



九州海軍航空隊

九州海軍航空隊（きゅうしゅうかいぐんこうくうたい）は、昭和19年7月10日に編成された最初の乙航空隊のひとつ。鹿屋飛行場を拠点とし、西日本各地の飛行場を管轄した。沖縄の戦いでの特攻作戦、以後の本土防衛防空作戦に多く関与している。練習航空隊の実施部隊化、練習基地の実施部隊転用などで規模も拡大し、昭和20年3月20日より「司令官」を隊長とした。20年春以降の乙航空隊細分化によって、管轄区は九州南部に縮小した。

主な航空基地

- 鹿屋飛行場・国分飛行場・出水飛行場・串良飛行場・笠之原飛行場・岩川飛行場

(a)「海軍乙航空隊」のスクリーンショット一部（ハイライトによる強調は著者によるもの）

トークン	真のラベル	予測ラベル		
		LayoutLM	BERT	BERT+100K
乙	O	O	O	O
航空	O	O	O	O
隊	O	O	O	O
の	O	O	O	O
ひと	O	O	O	O
つ	O	O	O	O
。	O	O	O	O
鹿屋	B-基地	O	B-基地	B-基地
飛行	I-基地	O	I-基地	I-基地
場	I-基地	O	I-基地	I-基地
を	O	O	O	O
拠点	O	O	O	O
と	O	O	O	O
し	O	O	O	O
、	O	O	O	O
		…		
主な	O	O	O	O
航空	O	O	O	O
基地	O	O	O	O
鹿屋	B-基地	B-基地	B-基地	B-基地
飛行	I-基地	I-基地	I-基地	I-基地
場	I-基地	I-基地	I-基地	I-基地
・	O	O	O	O

(b) ハイライト箇所に対する系列ラベリング結果

図 4: レイアウト情報が抽出性能にネガティブに影響を及ぼしたと考えられる例