

『日本人名辞典』からの歴史人物情報の抽出： Few-shot 学習による古文の固有表現抽出の試み

苑 広媛（立命館大学 情報理工学研究科）

李 康穎（日本学術振興会 特別研究員）

後藤 真（国立歴史民俗博物館）

木村 文則（尾道市立大学 経済情報学部）

前田 亮（立命館大学 情報理工学部）

概要：歴史研究を行うにあたり、人物に関する情報は重要な基盤情報となっている。自然言語処理技術を用いて辞書類から自動で人物属性の情報抽出を行うことにより、人文学研究を行う専門家のための支援を提供できると考えられる。一方で、機械学習手法を用いた歴史情報抽出のための学習データの作成にはコストがかかるという問題がある。そこで本研究では、『日本人名辞典』からの歴史人物情報抽出における few-shot 学習（few-shot learning: FSL）の活用を試みる。

キーワード：日本人名辞典, Few-shot 学習, 固有表現抽出

Extraction of information on historical figures from a Japanese Biographical Dictionary: An attempt to extract named entities from classical Japanese by few-shot learning

Guangyuan Yuan (Graduate School of Information Science and Engineering, Ritsumeikan University)

Kangying Li (JSPS Research Fellow)

Makoto Goto (National Museum of Japanese History)

Fuminori Kimura (Faculty of Economics Management and Information Science, Onomichi City University)

Akira Maeda (College of Information Science and Engineering, Ritsumeikan University)

Abstract: In current historical research, biographical information of historical figures is becoming important fundamental information. Automatic extraction of personal attributes from dictionaries and encyclopedias using natural language processing technology could provide support for experts in the study of humanities. However, preparing the training data for machine-learning-based information extraction from historical documents is costly. In this study, we try to use few-shot learning (FSL) in extracting the personal attributes of historical figures from the Japanese biographical dictionary.

Keywords: Japanese biographical dictionary, Few-shot learning, Named entity recognition

1. まえがき

『日本人名辞典』は国文学者芳賀矢一の著作であり、古代から明治まで五万余名の人名を収録し、簡潔に記述した人名辞典である。歴史研究を行うにあたり、人物情報は非常に重要な基盤情報である。人名辞典や人文学関連の辞書類からの情報抽出は、人文学研究の基盤を構築するだけでなく、機械学習手法を用いた情報抽出の研究における、学習データ作成のコストの問題を解決できる。しかし、既存の古典資料はまだ多くの基盤情報が不足している状況である。

安岡ら[1]は、MeCab による古典中国語の形態素解析のアプローチを提案しており、このための漢文コーパスを構築し、さらに固有表現抽出タスクにおいて、地名の抽出では高い精度を達

成したが、官職と人名の自動抽出にはまだ解決すべき問題が残されているとしている。相良ら[2]は、MeCab の現代語および古文の 6 種類の辞書を組み合わせ、江戸時代中期に書かれた『養生訓』の校訂および抄訳のテキストデータを対象とした形態素解析の実験を行った結果、「どのような文書も適切に語分割できる汎用的な形態素解析用の辞書は存在しない」ことを実証した。

従って、より少ないデータで学習可能なアルゴリズムが非常に重要となる。few-shot 学習（few-shot learning：以下 FSL）とは、比較的大量のデータを必要とするファインチューニングとは対照的に、非常に少量のデータを機械学習モデルに提示する手法を指す。学習データに基づき学習したモデルで、3～10個のラベル付き用例をサポートとして参照するだけで、他に同年

代のデータ量が少ないテキストでも自動ラベル付けや情報抽出を行うことができる。本研究では、『日本人名辞典』からの歴史人物情報抽出における FSL の活用を試みる。図 1 に、本研究で使用する芳賀矢一著『日本人名辞典』の原文の一部を示す。また、本研究で抽出する人物情報の例として、人物「阿蘇惟豊」の解説文と、そこから抽出すべき属性を図 2 に示す。

本研究では、第 3 著者の後藤により作成された、『日本人名辞典』の全解説文のテキストおよび、その一部について手作業で人物に関する属性が付与されたものを実験データとして用いる。



図 1 『日本人名辞典』(出典：国立国会図書館，請求記号 281.03-H117n)

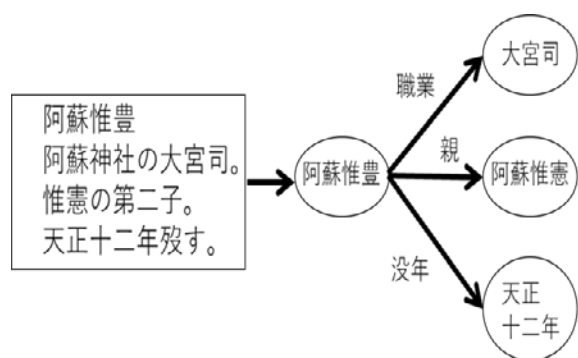


図 2 人物「阿蘇惟豊」の解説文と、そこから抽出すべき属性の例

2. 関連研究

本研究では、デジタルテキスト化された人名辞典の解説文から人物の属性情報を抽出するための情報抽出手法として、固有表現抽出を使用する。固有表現抽出 (named entity recognition: 以下 NER) とは、人名や組織名などの固有名詞や日付表現、時間表現など、特定実体を指す固有表現 (named entity) をテキストから自動的に抽出する技術である。辞書類から知識の自動抽出を行うための手段として、パターンベースの

手法や既存の NER ツールが存在する[3]。近年、深層学習が注目されるようになり、深層学習を用いたアプローチの研究が多く報告されている。白井ら[4]は、本研究の対象と同じ『日本人名辞典』に対して、BiLSTM-CRF モデルを用いて人物の属性を抽出する研究を行い、全体的に高い精度が得られ、特定のパターンで記述されている属性に対して特に高い抽出精度が得られることを示した。川端ら[5]は、[4]と同様に BiLSTM-CRF モデルを用いて、歌舞伎役者の芸評書である役者評判記から役者の情報を抽出する研究を行い、役者に関する 5 種類の属性について、F 値の平均で 90%以上の精度が得られることを示した。

しかし、一般的な深層学習モデルは大量の学習データが必要であり、少ない学習データでは適切に学習できないことが多い。FSL は、事前知識を用いることで、教師あり情報で少数のサンプルしか含まない新しいタスクにも汎化でき、学習データ不足の問題を解決することができる。Ma ら[6]は、prompt learning に基づく NER モデルの認識効率の問題を解決するために、Entity-oriented LM (EntLM) のチューニング法を提案した。Cui ら[7]は、生成モデルアプローチにおける NER の問題を解決するために、Template NER の学習段階でテンプレートを構築し、Generative BART モデルを系列ラベリングのタスクに適用した。Yao ら[8]は、Universal Information Extraction (UIE) を提案し、一般情報抽出モデルを訓練してオープンソースで提供した。

現状では、このような few-shot NER を古文に応用した研究は少ない。1 章で述べたように、学習データ作成のコストが高い古文の NER に対して、FSL の技術は非常に有用であると考えられる。本研究では、FSL に基づく固有表現抽出器を利用する。

3. 提案手法

本研究では、固有表現抽出手法として Example-based NER モデル[9]に基づき、青空文庫 DeBERTa モデルに基づく NER モデルを使用する。本研究で使用するモデルを図 3 に示す。

本研究のモデルはサポート (固有表現のラベルを含む文) のサンプル数が少ないことを想定しているため、すべての属性の種類に対応する固有表現はわずかしかない。クエリ (固有表現抽出の対象である、ラベル付けされていないテキスト) が与えられると、モデルはサポート内の固有表現の種類に従って、クエリに対応する固有表現を探す。つまり、本モデルは、固有表

現の範囲を識別する部分と、これに基づいてデータを分類する部分の二つに分けられる。

まず一つ目の部分では、サンプルの文内から、固有表現である可能性が最も高い範囲の開始位置と終了位置を識別する。具体的な方法は、クエリごとに、対応するサポートが構築され、ある固有表現の範囲を示すタグとして $\langle e \rangle$ や $\langle /e \rangle$ などの特殊文字が生成され、サポート内の固有表現の前後に追加される。BERT を使用してクエリとサポートをそれぞれエンコードし、サポート内の $\langle e \rangle$ と $\langle /e \rangle$ とクエリ内の各文字単位のトークンの一致する位置を計算し、クエリ内の固有表現の範囲の開始位置と終了位置を見つける。

モデルの二つ目の部分では、最初のステップで識別された固有表現の範囲に基づいて、この固有表現の範囲がどの固有表現に対応するかをさらに識別する。このステップの具体的な方法は最初のステップと似ており、サポートサンプルを使用し、クエリ内の各トークンが各固有表現の種類の開始位置と終了位置である確率を計算する。

図 3 における二つの BERT は、クエリ Encoder とサポート Encoder として用いられる。サポート Encoder は固有表現の出現位置の計算とカテゴリ推測のために、ラベル付きデータから分散表現を抽出する。クエリ Encoder は、ラベルなしの解説文から分散表現を抽出する。抽出されたクエリとサポートベクトルは、文レベルのアテンションを組み込んだトークンレベルの類似度の計算により固有表現の分散表現を学習する。学習済みモデルは、新しく与えられた少ないラベル付きテキストのサポート例から、クエリ文書の固有表現の情報を予測できる。

4. 実験

4.1 既存の現代語 NER モデルによる実験

本研究では、まず既存の現代語 NER モデルを用いて、『日本人名辞典』の解説文テキストから人物情報を抽出する実験を行った。

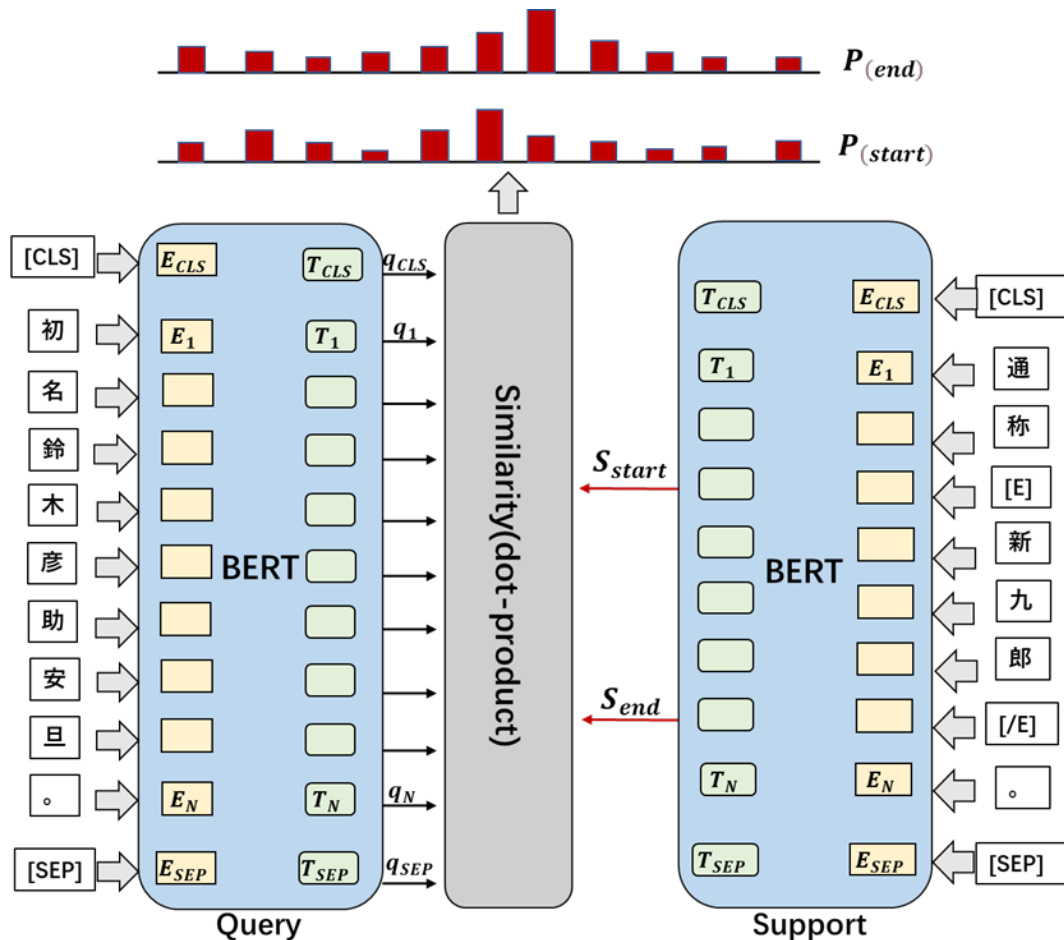


図 3 本研究で用いる Example-based NER モデル



図4 ner_ud_gsd_glove_840B_300d モデルによる抽出結果の例

一つ目の実験で用いた ner_ud_gsd_glove_840B_300d モデルは、Apache Spark ML 上に構築された自然言語処理ライブラリである Spark NLP に含まれる言語モデルの一つである[10]。このモデルは Universal Dependencies (UD) で訓練され、テキスト内の人物、地名、組織名などの固有表現を抽出するために使用できる。抽出結果の例を図4に示す。認識される固有表現は、出来事 (EVENT)、政府を持つ地域名 (GPE: geo-political entity)、運動 (MOVEMENT)、人名 (PERSON) の四つがある。しかし、図4の例では、たとえば動詞である「撰す」を人名と誤認識し、地域名である「日本」を出来事と誤認識するなど、非常に不正確であることがわかる。

上記の実験結果より、文語体である『日本人辞典』の解説文に対して現代語 NER モデルを用いた場合の抽出精度は非常に低く、文語体の固有表現抽出には不適切であることがわかる。

最も簡単な NER の方法としては、辞書や形態素解析結果、正規表現などに基づくルールを用いて、単語列にラベリングする方法がある。ルールベースの手法を利用する可能性を検討するため、二つ目の実験を行った。

本検証実験に用いた ja_core_news_sm は、固有表現認識システム spaCy v2.0 に含まれる日本語のモデルである[11]。サブワード特徴と「ブルーム」埋め込みを用いた単語埋め込み戦略、残差接続を用いた深い畳み込みニューラルネットワーク (CNN) および固有表現のための新しい遷移ベースのアプローチを特徴とする。実験結果を表1に示す。動詞である「号す」は「NOUN (名詞)」と誤認識されている。また、「天保十四年六十八年九月六日没。」の部分为例にとると、本来 NUM (数値) であるべき前半部分の「十四」が「PROPN (固有名詞)」と誤認識され、後半部分の「六十八」が「NOUN (名詞)」と誤認識されている。

表1 ja_core_news_sm モデルによる抽出結果の例

曇亀	PROPN nmod
、	PUNCT punct
又	CCONJ cc
拙	PROPN compound
斎	NOUN nmod
と	ADP case
号す	NOUN ROOT
。	PUNCT punct
天保	PROPN compound
十四	PROPN compound
年	NOUN compound
九	NUM compound
月	NOUN compound
六	NUM compound
日	NOUN compound
没	NOUN ROOT
。	PUNCT punct
年	NOUN compound
六十八	NOUN ROOT
。	PUNCT punct

上記のように、現代語の固有表現抽出器および形態素解析器による実験を行った結果、抽出精度は低く、現代語の形態素解析器の結果をルールベースの NER 手法に応用することも難しいと考えられる。

4.2 データ分析

1章で述べた、専門家により人物に関する属性が付与された解説文のうち、1,717 件を対象として分析を行った。抽出対象の属性として、「地理情報」、「別名」、「所属組織」、「時代」、「著作」、「作歌」、「親」、「妻/夫」、「仕えた人」、「死没年月日」、「死没時齢」の12種類の属性がある。本研究では、これらを全人物に共通の属性（「地理情報」、「別名」、「所属組織」、「時代」、「著作」、「作歌」）と、各人物と一对一の関係を持つ属性（「親」、「妻/夫」、「子供」、「仕えた人」、「死没年月日」、「死没時齢」）の2種類に分けて扱うこととする。

「地理情報」、「別名」、「所属組織」、「時代」、「著作」、「作歌」を共通属性とする理由は、これらの属性はその特性を客観的に特徴付ける必要があるためである。また、「親」、「妻/夫」、「子供」、「仕えた人」、「死没年月日」、「死没時齢」を一对一属性として人物ごとの解説文に分ける理由は、これらの属性は関係情報を持つためである。例えば、解説文の中で人物の名前が「子供」属性としてラベル付けされて

いる場合、これは「別名」属性にも適用される可能性があり、「別名」の固有表現の予測の精度に影響を与える可能性がある。つまり、人物間の関係を表すラベルによる関係抽出のタスクと、固有表現ラベルによる固有表現抽出のタスクを混在して評価するのは無理があると考ええる。したがって、一対一属性の抽出を共通属性の抽出とは別のタスクとして扱うこととする。

表 2, 表 3, 表 4 に、ラベルを与えたデータ数を示す。ここでの「マッチング」は、テキストにラベルを付与する作業を示す。元データは専門家が各人物の解説文に対して属性をラベル付けしたものであり、人物ごとに付けられたラベルをそのまま実験に用いるのが、ここでの「人物ごとの解説文マッチング」である。各人物の解説文に登場する地名、人名などの共通属性を全人物の解説文に適用し属性にラベルを与えるのが、ここでの「全解説文マッチング」である。共通属性に対しては、人物ごとの解説文と全人物の解説文のマッチングを行い、重複項目を削除する。

表 2 共通属性の人物ごとの解説文マッチング

(共通) 属性	人物ごとの解説文
地理情報	167
別名	1730
所属組織	300
時代	70
著作	637
作歌	3

表 3 共通属性の全解説文マッチング

(共通) 属性	全人物の解説文
地理情報	273
別名	1798
所属組織	473
時代	76
著作	647
作歌	13

表 4 一対一のマッチング

(一対一) 属性	人物ごとの解説文
親	314
妻/夫	39
子供	22
仕えた人	113
死没年月日	844
死没時齢	651

4.3 実験概要と実験結果

前処理済みのデータを用いて提案手法の実験を行った。学習データは、前節で述べた 1,717 件

であり、そのうち 20% をテストデータとして使う。実験結果を表 5 に示す。共通属性に対して、人物ごとの解説文と全人物の解説文の正解率は、それぞれ 41.7% および 35% となった。属性によって精度にばらつきが存在するものの、少量の学習データを使用する NER タスクとしては、全体的に比較的高い抽出精度を実現している。特に、一対一属性はロスが最も小さく、正解率も 62% と、FSL に基づく NER としては比較的高い結果が得られた。

表 5 提案手法の実験結果

	(共通) 人物ごとの解説文	(共通) 全人物の解説文	(一対一) 人物ごとの解説文
Loss	0.00361	0.00456	0.00140
Val loss epoch	0.00500	0.00533	0.00384
Val acc epoch	0.41700	0.35000	0.62000
Train loss epoch	0.00415	0.00501	0.00292

実験の結果の一例として、属性「親」の抽出の結果を図 5 に示す。属性「親」のテストデータと学習データは合計で 314 件であり、そのうち 20% (約 60 件) をテストデータに使用し、80% (約 250 件) を学習データに使用した。本実験では複数回のサンプリングを行い、毎回ランダムにテストデータから 5 件、学習データから 10 件を抽出した。3 回のサンプリングで得られる結果の平均精度は 100 % 近くになった。図 5 における「start」は開始位置、「end」は終了位置、「entity_value」は固有表現、「parent」は予測属性、「score」は確度を表す。

女歌人。平兼盛 parent の女、赤染時用に養はる。上東門院に仕へ、後大江匡衡に嫁す。拾遺、後拾遺、詞花、千載、新古今、新勅、続後撰、続古今、続拾遺、玉葉、続千載、続後拾遺、風雅、新千載、新拾遺、新後拾遺の作家。

```
[34]: output
[34]: [{"start": 4,
      'end': 7,
      'entity_value': '平兼盛',
      'score': 1.0,
      'label': 'parent'}],
[{'start': 13,
  'end': 17,
  'entity_value': '安藤親重',
  'score': 1.0,
  'label': 'parent'}],
[{'start': 11,
  'end': 15,
  'entity_value': '戸田忠之',
  'score': 1.0,
  'label': 'parent'}],
[{'start': 0,
  'end': 4,
  'entity_value': '桓武天皇',
  'score': 1.0,
  'label': 'parent'}],
[{'start': 5,
  'end': 9,
  'entity_value': '三条天皇',
  'score': 1.0,
  'label': 'parent'}]]
```

図 5 属性「親」によるテキストの予測例

4.4 対照実験

本実験で使用したモデルの性能を評価するため、サポートセットのサンプル数が異なる設定の下での学習能力をテストした。本実験では、4.2 節で説明したデータセットからサポートサンプルを使用し、推論に使用するサポートセットを、全体から固有表現タイプごとに一定数の例をランダムにサンプリングした。

表 6 提案手法の対照実験の結果

サポート サンプル 数	正解率 (Val acc epoch)		
	(共通) 人物ごと の解説文	(共通) 全人物の 解説文	(一対一) 人物ごとの 解説文
5 shot	0.185	0.133	0.362
10 shot	0.210	0.189	0.230
15 shot	0.216	0.209	0.167
20 shot	0.250	0.244	0.131

サポートサンプル数を 5 shot, 10 shot, 15 shot, 20 shot の範囲で変化させた場合に、共通属性および一対一属性の正解率の変化を確認する実験を行った結果を表 6 に示す。一対一属性の場合、サンプル数が増えると正解率は 36.2% から 13.1% に低下した。これは、共通属性よりも、固有表現が出現する文脈においてより多くの表現が存在するためであると考えられる。サポートサンプルの数が増えると、コンテキストの構造に対するモデルの信頼度が低下し、データがモデルの予測能力を減少させる。これが、「親」、「妻/夫」、「仕えた人」、「死没年月日」、「死没時齢」を一対一属性のみとして分類する理由である。このうち共通属性における人物ごとの解説文と全人物の解説文に対しては、サポートサンプル数の増加に伴い、正解率が高くなった。その中で最も精度が向上したのは共通属性における全人物の解説文で、正解率は 13.3% から 24.4% に向上した。

FSL による NER の精度の現状については、Example-NER の論文[9]でも説明されているように、現状では精度が全体的に高くないが、古文における固有表現抽出に新たな研究手法の可能性を示すことできたと考える。また、本手法の対照実験より、モデルの有効性が向上し、モデルの汎化能力が向上している結果を得られた。

5. あとがき

本研究では、少ない学習データで利用可能な FSL を利用して、『日本人名辞典』から人物に関する属性の抽出を行う手法を提案し、実験を行

った。また、モデルの性能評価のための対照実験を行った。実験結果から、FSL に基づいて少ないデータで学習したモデルが古文においても現代文と同等の精度が得られ、モデル性能も良好であることを示した。

今後は、既存手法との比較を含む大規模な評価実験や、抽出結果を利用した人物関係の抽出などを検討する。

謝辞

本研究の一部は JSPS 科研費 20K12567 の助成を受けたものである。

参考文献

- [1] 安岡孝一, ウィッテルン クリスティアン, 守岡知彦, 池田巧, 山崎直樹, 二階堂善弘, 鈴木慎吾, 師茂樹: 古典中国語(漢文)の形態素解析とその応用, 情報処理学会論文誌, Vol. 59, No. 2, pp. 323-331, 2018.
- [2] 相良かおる: 『養生訓』の自動形態素解析における辞書の影響, じんもんこん 2018 論文集, pp. 153-160, 2018.
- [3] Samet Atdag and Vincent Labatut: A comparison of named entity recognition tools applied to biographical texts, Proc. ICSCS 2013, pp. 228-233, 2013.
- [4] 白井圭佑, 松崎真里, 森信介, 後藤真: 人名辞典からの知識抽出, 情報処理学会論文誌, Vol. 63, No. 2, pp. 293-301, 2022.
- [5] 川端恵大, 前田亮, 赤間亮: 役者評判記を用いた役者情報の抽出, じんもんこん 2021 論文集, pp. 200-205, 2021.
- [6] Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang: Template-free Prompt Tuning for Few-shot NER, Proc. NAACL 2022, pp. 5721-5732, 2022.
- [7] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang: Template-Based Named Entity Recognition Using BART, Proc. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1835-1845, 2021.
- [8] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu: Unified Structure Generation for Universal Information Extraction, Proc. ACL 2022, pp. 5755-5772, 2022.
- [9] Morteza Ziyadi, Yuting Sun, Abhishek Goswami, Jade Huang, and Weizhu Chen: Example-Based Named Entity Recognition, arXiv:2008.10570 [cs.CL], 2020.
- [10] Veysel Kocaman and David Talby: Spark NLP: Natural Language Understanding at Scale, Software Impacts, Vol. 8, 2021.
- [11] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum: Fast and Accurate Entity Recognition with Iterated Dilated Convolutions, Proc. EMNLP 2017, pp. 2670-2680, 2017.