クラス定義文を用いた Wikipedia 記事の分類

林歓樹 [1] 中山功太 [2,3] 関根聡 [2]
[1] 早稲田大学基幹理工学部情報理工学科 [2] 理化学研究所 AIP [3] 筑波大学情報生命学術院
rainline@akane.waseda.jp {kouta.nakayama, satoshi.sekine}@riken.jp

概要

本稿では、文書分類の一手法としてクラス定義文と深層学習を用いた手法を提案する。本手法は、Zero-shot 設定にも適用可能といった特徴がある。提案手法のモデルを Wikipedia 構造化プロジェクト「森羅 2021」の多言語分類タスクで提出した結果、全参加者のモデルの中で 30 言語中 17 言語でトップの性能を示した。また、Zero-shot 設定でも実験を行い、提案手法が一定の有効性を持つことを示した。その結果から Zero-shot 予測に適した定義文について議論を行なった。

1 はじめに

近年深層学習を文章分類タスクへ適用した事例が多く報告されている。その中でも訓練時に現れなかったクラスを分類時に認識できるようにする Zero-shot 学習が盛んに研究されている [1]。

「森羅プロジェクト」[2] では、Wikipedia 上の知識を機械が認識できるような形式に構造化することを目標に、多くの人の協働のもとでリソースの構築を行なっている。その中の分類タスクでは、Wikipedia 記事を拡張固有表現[1)] に基づいたカテゴリーへ分類する。

本稿では、Wikipedia 記事の分類タスクを解くため、クラス定義文を深層学習と併せて利用することで分類を行う手法を提案する。実験によって、定義文を用いたアプローチが文章分類タスクで有効であること、また、Zero-shot 設定においても定義文を用いたアプローチが有効であることを示す。

2 SHINRA2021ML

SHINRA2021ML は、30 言語の Wikipedia 記事を拡張固有表現に基づいた約 220 のカテゴリーに分類をするタスクである。本タスクでは分類済み日本語 Wikipedia 記事から言語間リンクが張られている他言語の記事を訓練データとして用いている。本来は複数クラスへの分類タスクだが、訓練データのうち複数クラスに分類される例が 2.50%と非常に少ないため、今回は単一クラスの分類タスクとして扱う。

3 関連研究

今まで提案された多言語における Wikipedia 記事の分類手法として、多言語 BERT モデル [3] を全言語で学習させたのち、各言語のモデルを作成してその言語ごとの学習データのみで fine-tuning するという手法が提案されている [4]。その際用いられる学習データは Wikipedia 記事本文のみである。また本文以外の情報を学習データに利用した手法も提案されている [5]。記事本文の他に、記事同士の関係を表した知識グラフ、画像情報、記事のレイアウト、拡張固有表現の階層情報を学習データとして利用している。また、Wikipedia 自身のカテゴリ情報のみを用いて分類する手法も提案されている [6]。

Aly らは、クラス定義文を利用した Zero-shot 固有表現抽出の手法を提案している [1]。対象とする文にクラス定義文を連結して深層学習モデルに入力することで訓練時に現れないクラスの抽出を行なっている。

本研究では、Wikipedia の記事の分類にクラス定義文の情報を利用する手法を提案する。さらに Zero-shot 設定での分類も行う。

4 提案手法

一般的なクラス分類モデルでは、深層学習モデルの出力ベクトルと、各行が各クラスに対応する学習可能な行列との積を取ることでクラス予測を行う。そのため、各クラスに付与された情報を用いることができない。対して、提案手法では後者の行列をクラスに付与された定義文を元に生成する。つまり、文章を埋め込むモデルの出力ベクトルとクラス定義文を埋め込むモデルの出力行列の積により最終的なクラスを予測する。

1） http://liat-aip.sakura.ne.jp/ene/ene8.1/definition_jp/

Wikipedia 記事の埋め込み表現の取得方法　初めに Wikipedia 記事の文章を、WordPiece[7] により分割し、トークン列 $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ を得る。p は深層学習モデルへの入力長である。WordPiece はサブワードの考え方を用いた分割の一手法である。次に、104 言語で事前学習された多言語 BERT モデル [3] を用いて得られたトークン列から埋め込みベクトル $\mathbf{v} \in \mathbb{R}^n$ を得る。[2] ここで n は埋め込みベクトルの次元数である。

クラス定義文の埋め込み表現の取得方法　SHINRA2021ML タスクでは階層的な拡張固有表現の末端カテゴリーへの分類を行う。そのため我々は本実験では、末端カテゴリーのクラス定義文のみを用いる。拡張固有表現の i 番目の末端カテゴリの定義文に対し、Wikipedia 記事と同様の前処理を行い、トークン列 $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,q})$ を得る。q は深層学習モデルへの入力長である。その後、学習済み多言語 BERT を用いて、埋め込みベクトル $\mathbf{g}_i \in \mathbb{R}^n$ を得る。また、すべての定義文を埋め込むことで行列 $G \in \mathbb{R}^{c \times n}$ を得る。c は末端カテゴリーの総数である。

クラス分類　得られた文章埋め込みベクトル v と定義文埋め込み行列 G の積により、最終的な各クラスの予測確率 $\mathbf{p} \in \mathbb{R}^c$ を得る。

$$\mathbf{s} = \mathrm{softmax}(\mathbf{t}G^\top)$$

最終的な予測クラス y は、次のように求まる。

$$y = \underset{i \in \{1,\ldots,c\}}{\arg\max}\ \mathbf{s}_i \tag{1}$$

損失関数には、交差エントロピー誤差を用いる。

5　実験

5.1　実験 1: ベースラインとの比較

データセット　森羅 2021 において配布された 30 言語の Wikipedia 記事とその分類先カテゴリの教師データを学習に用いた。評価は森羅 2021 のテストデータで行った。

実験対象モデル　以下の 2 つのモデルの評価を行なった。

mBERT　ベースラインモデル。multilingual BERT によって Wikipedia 記事の本文の埋め込み表現を得て、1 層の線形層を用いて分類を行うモデル。

mBERT_def　提案手法。2 つの multilingual BERT を用いて Wikipedia 記事と定義文の埋め込み表現をそれぞれ得て、その類似度によって分類を行うモデル。

学習設定　分類先カテゴリ数は $c = 221$ である。記事と定義文の埋め込みベクトルの次元数は $n = 768$ を使用した。各モデルはエポック数 8、ミニバッチサイズ 64 で学習を行なった。学習率は 5.0×10^{-5} に設定した。Wikipedia 記事の埋め込み表現の入力長は $p = 512$ に設定した。また、定義文の埋め込み表現の入力長は $q = 93$ に設定した。

5.2　実験 2: Zero-shot 設定

Zero-shot 設定での実験を行なった。

データセット　森羅 2021 において配布された 30 言語の Wikipedia 記事とその分類先カテゴリの教師データのうち、正解クラスを出現頻度順にソートしたときに、上位 80 %のクラスを学習データ、80〜90 %のクラスをテストデータとして用いた。評価はテストデータを用いて行なった。クラスの分割については表 1 に示す。

実験対象モデル　mBERT_def のみで実験を行なった。

学習設定　分類先カテゴリ数は学習時は $c = 32$, 推論時は $c = 26$ である。記事と定義文の埋め込みベクトルの次元数は $n = 768$ を使用した。各モデルはエポック数 8、ミニバッチサイズ 64 で学習を行なった。学習率は 5.0×10^{-5} に設定した。Wikipedia 記事の埋め込み表現の入力長は $p = 512$ に設定した。また、定義文の埋め込み表現の入力長は $q = 93$ に設定した。

6　評価結果

6.1　実験 1

各モデルの森羅 2021 のテストデータによる各言語に対する評価結果を表 2 に示す。mBERT_def の結果は、mBERT より 28 言語のうち 17 言語で性能が向上した。

また、全参加者のモデルの中で、mBERT_def が 30 言語のうち 17 言語で最も良い性能を示した。定義文を用いたアプローチが性能の向上につながっていることがわかる。

2)　この際トークン列に対してそれぞれ文章の先頭と終端を示す特殊トークン [CLS] と [SEP] を結合する。

表 1 実験 2 のクラス分割

	クラス名 (出現数)
train	Person(1305318), CONCEPT(460272), City(437884), Province(181974), Movie(165102), Doctrine_Method_Other(123633), Competition(116828), Music(107026), Sports_Team(93168), Date(89343), IGNORED(76758), Company(76015), Weapon(73026), Book(67854), Product_Other(54098), Station(53163), Broadcast_Program(50421), Compound(48293), Show_Organization(48238), Planet(42883), Software(41887), Digital_Game(41780), Character(39136), Ship(36995), Flora(35272), Bird(33463), Car(32478), Position_Vocation(30681), War(29286), Mammal(28533), Aircraft(27814), Island(27422)
test	Animal_Disease(27304), Country(23910), School(23258), Animal_Part(23011), Sports_Facility(21884), Academic(21735), Astronomical_Object_Other(21678), Dish(19490), Military(18382), Theory(18132), Airport(17862), Award(17538), River(17252), Mountain(16783), Political_Party(16406), Railroad(16218), Era(16053), Spaceship(16049), Archaeological_Place_Other(15968), Language_Other(15943), Food_Other(15650), Domestic_Region(15304), Reptile(14914), Ethnic_Group_Other(14817), Worship_Place(14690), Government(14506), God(14362), Facility_Other(14066)

表 2 実験 1 のテストデータに対する結果

言語	mBERT	mBERT_def	差
Czech	78.67	81.70	3.03
Romanian	89.77	92.79	3.01
Chinese	83.35	85.59	2.23
Norwegian	83.39	85.06	1.67
Dutch	83.56	85.11	1.55
Catalan	80.26	81.80	1.54
Danish	78.51	79.92	1.41
Finnish	83.06	84.30	1.24
Korean	78.01	79.05	1.04
German	76.88	77.89	1.01
Spanish	82.01	82.87	0.86
Ukrainian	82.51	83.32	0.81
English	83.00	83.42	0.42
Russian	79.80	80.20	0.41
Hungarian	89.93	89.93	0.00
Bulgarian	84.58	84.58	0.00
Thai	79.92	79.92	0.00
Hindi	86.35	86.14	-0.20
Italian	82.65	82.45	-0.20
Arabic	89.26	89.05	-0.21
Polish	85.93	85.71	-0.22
Turkish	83.50	83.10	-0.40
French	83.39	82.96	-0.42
Hebrew	80.32	79.72	-0.61
Swedish	82.96	82.08	-0.88
Vietnamese	90.28	88.19	-2.08
Persian	85.27	82.78	-2.49
Indonesian	86.77	82.95	-3.82
平均	83.35	83.66	0.31

6.2 実験 2

Zero-shot 設定での各クラスごとの評価結果を表 3 に示す。全体としての F1 値のマイクロ平均は 0.30 とあまり高くない数字だが、クラスごとに見ると 0.70 を超えるクラスがいくつかあり、定義文を用いたアプローチが一定の有効性を示している。

予測性能の差に関する定義文の分析 最も F1 値が高かった 3 クラスと低かった 3 クラスについて詳細な分析を行なった結果を表 4 に示す。ここで本文とは Wikipedia 記事本文の先頭 512 単語のことである。また、定義文の単語とは、定義分に含まれる単語から"an"や"of"などのストップワードを除いたものである。F1 値が高い 3 クラスのうち Reptile と Era は、定義文に含まれる単語のうち本文に含まれる単語の割合が高く、F1 値が 0 の 3 クラスは低いことがわかる。このことから、定義文に含まれる単語が Wikipedia 記事本文に含まれていると F1 値が高くなる傾向があると言える。しかし、Astronominal_Object_Other クラスに関しては、定義文の単語のうち本文に含まれる単語の割合が低いのにもかかわらず高い F1 値を出している。これは表 5 に示した定義文を見ると、Astronominal_Object_Other クラスの定義文は例示を多用していたり注釈がついていたりと他の定義文と比べて長くなっている。これによってより多い単語を定義文に含むため、Wikipedia 記事本文に定義文の単語が含まれる割合が高くなると考えられる。実際、表 4 に示したように、Astronominal_Object_Other クラスでは、本文に定義文の単語が一つでも含まれる割合が高くなっている。以上から、定義文の単語が記事本文に含まれる割合が高いとき、または定義文の長さが長く、定義文の単語が記事本文に含まれやすくなっているときに F1 値が高くなる傾向があると考えられる。

表 3 実験 2 のテストデータに対する結果

クラス名	precision	recall	F1
Reptile	0.88	0.99	0.93
Era	0.93	0.62	0.74
Astronomical_Object_Other	0.71	0.72	0.71
Award	0.55	0.83	0.67
Military	0.97	0.49	0.65
Domestic_Region	0.56	0.68	0.61
Political_Party	0.4	0.74	0.52
Railroad	0.83	0.3	0.44
School	0.35	0.25	0.29
Ethnic_Group_Other	0.22	0.33	0.27
Language_Other	0.14	0.92	0.25
Dish	0.15	0.54	0.24
Facility_Other	0.16	0.41	0.22
Mountain	0.26	0.14	0.19
Worship_Place	0.4	0.12	0.19
Government	0.18	0.17	0.17
Archaeological_Place_Other	0.09	0.42	0.15
Airport	0.53	0.05	0.1
Food_Other	0.08	0.12	0.09
Animal_Part	0.67	0.03	0.06
Sports_Facility	0.72	0.03	0.06
Spaceship	0.97	0.03	0.05
River	0.08	0.02	0.03
God	0.74	0.01	0.02
Country	0.01	0	0.01
Animal_Disease	0	0	0
Academic	0	0	0
Theory	0.02	0	0
マイクロ平均			0.30
マクロ平均	0.41	0.30	0.27

表 4 実験 2 の定義文に関する分析

クラス名	F1	定義文の単語のうち本文に含まれる単語の割合	本文に定義文の単語が一つでも含まれている割合
Reptile	0.93	0.503	0.848
Era	0.74	0.426	0.844
Astronomical_Object_Other	0.71	0.122	0.962
animal_disease	0	0.267	0.768
Academic	0	0.296	0.674
Theory	0	0.306	0.817

表 5 定義文の例

クラス名	定義文
Reptile	A name of a reptile.
Era	An expression of an era.
Astronomical_Object_Other	A name of an astronomical object (1.5.4 Astronomical_Object) that do not belong to any of the other subordinate categories (1.5.4.1 - 1.5.4.3).Examples are a galaxy, nebula, comet, satellite, interstellar substance, planetary system, etc. An artificial satellite is not included here but in 1.7.17.4 Spaceship Category.
animal_disease	A name of an animal disease or injury.
Academic	A name of an academic field.
Theory	A name of a theory or natural law.

7 おわりに

本研究では、森羅 2021 の多言語分類タスクにおいて、定義文を用いたアプローチが有効であること、また、Zero-shot の設定でも定義文を用いたアプローチがある程度有効であることを示した。また、Zero-shot 設定の実験から、どのようなクラス定義文が文書分類に役立つのかを示した。この結果から、Wikipedia の文書分類に限らない、他のタスクへの応用が考えられる。また、各言語ごとの詳細な分析を行うことで、更なる性能向上につなげられる可能性がある。それらについては今後の課題とする。

謝辞

本研究は JSPS 科研費 JP20269633 の助成を受けたものです。

参考文献

[1] Rami Aly, Andreas Vlachos, and Ryan McDonald. Leveraging type descriptions for zero-shot named entity recognition and classification. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics**, 2021.

[2] Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. Overview of shinra2020-ml task. In **Proceedings of NTCIR-15**, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **NAACL HLT 2019**, 2019.

[4] The Viet Bui and Phuong Le-Hong. Cross-lingual extended named entity classification of wikipedia articles. In **Proceedings of NTCIR-15**, 2020.

[5] Hiyori Yoshikawa, Chunpeng Ma, Aili Shen, Qian Sun, Chenbang Huang, Guillaume Pelat, Akiva Miura, Daniel Beck, Timothy Baldwin, and Tomoya Iwakura. Uom-fj at the ntcir-15 shinra2020-ml task. In **Proceedings of NTCIR-15**, 2020.

[6] Masaharu Yoshioka and Yoshiaki Koitabashi. Hukb at shinra2020-ml task. In **Proceedings of NTCIR-15**, 2020.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Moham-

mad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.