日本語大規模言語モデルにおける知識グラフを活用した意味理解性能の向上

中本裕大 [1]　瀬在恭介 [1]　元川凱喜 [1]　麻生英樹 [2]　岡崎直観 [3]
[1]SCSK 株式会社　[2] 産業技術総合研究所　[3] 東京工業大学
{yu.nakamoto,sezai,k.motokawa}@scsk.jp　h.asoh@aist.go.jp　okazaki@c.titech.ac.jp

概要

事前学習後の BERT は一般的な言語表現を獲得するが, ドメイン固有の知識が不足している. ドメイン固有の知識を BERT に注入する手法として, 知識グラフを活用した研究が行われている. 本研究では, 日本語 BERT モデルのファインチューニング時に知識グラフに記載されている知識を活用することによる, 下流タスクへの性能向上の可能性を検討した. また, LIME を用いて, 文章分類タスクにおいて, 知識を追加したことによる予測結果への影響調査を行った.

1　はじめに

近年, BERT [1] をはじめとする大規模言語モデルが言語処理の幅広いタスクにおいて高い精度を達成している. BERT は大規模なコーパスを用いて文脈情報を加味した単語表現を獲得することを可能とする. 事前学習後の BERT は一般的な言語表現を獲得するが, ドメイン固有の知識が不足している.

ドメイン固有にモデルを最適化する手法の 1 つとして, 事前学習またはファインチューニングの段階で外部の知識グラフ (以下 KG) から BERT に知識を注入する研究が行われている [2, 3, 4]. KG を BERT の学習に用いることで, テキストに書かれていないドメイン固有の知識などを考慮した言語表現の獲得が可能となる. BERT のファインチューニング時に KG の知識を活用する KI-BERT [4] では, GLUE [5] の 8 つのタスクで BERT と比較し著しい精度向上を達成した. 英語や中国語など日本語以外の言語では KG と BERT を掛け合わせた研究が活発に行われているのに対して, 日本語 BERT モデルを用いた手法の提案はまだ少ない. BERT のビジネス活用では, 計算資源や学習時間などの学習コストが課題となる. したがって, 知識グラフで表現された知識をファインチューニングで活用する手法は, ビジネス活用において需要が高い.

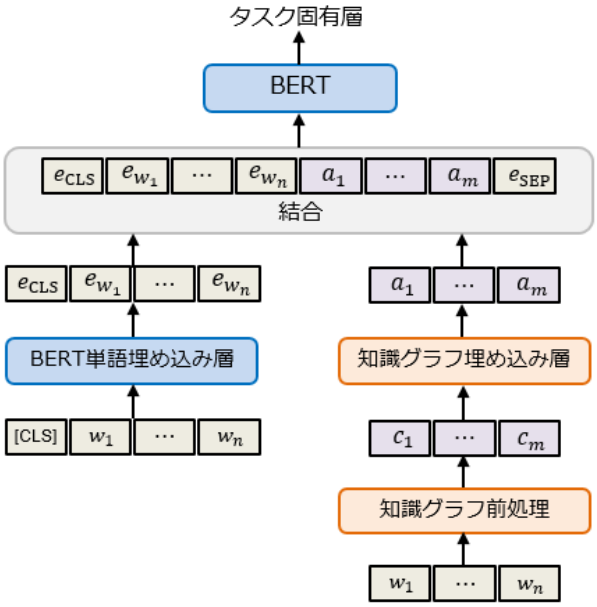


図 1: 提案手法の概要

そこで, 本研究では日本語 BERT モデルのファインチューニング時に KG の知識を活用することによる下流タスクの性能向上の可能性を検討した. JGLUE [6] の 5 つのタスクのうち 4 つのタスクで精度向上を確認した. また, 実際のビジネスで利用されるデータの一例として, 子育てオープンデータ協議会より公開されている子育て FAQ データでも実験を行い, 効果を確認した.

2　提案手法

本節では, 日本語 BERT モデルにおいて KG の知識を活用したファインチューニングを可能とするモデルを提案する.

提案するモデルの概要を図 1 に示す. 提案手法では, 入力として n 単語の単語列 $X = (w_1, ..., w_n)$ が入力されると, 入力に関連する知識群 $C = (c_1, ..., c_m)$ を**知識グラフ前処理**で取得する. ここで, c_i は知識

グラフのエンティティ (ノード) を指し, m は知識グラフ前処理により取得したエンティティ数となる. 本稿では, 知識グラフの前処理により獲得した c_i を「知識」と表現する. 取得した C は**知識グラフ埋め込み層**で知識埋め込み表現 $[a_1, ..., a_m] \in \mathbb{R}^d$ に変換される (d は次元数). BERT 埋め込み層により変換された単語埋め込み表現 $[e_{\text{CLS}}, e_{w_1}, ..., e_{w_n}] \in \mathbb{R}^d$ と知識グラフ埋め込み層で変換された知識埋め込み表現の結合を BERT に与えることで, KG の知識も考慮したアテンション計算を行う.

以降では, 日本語 KG の知識を BERT 内部で利用するにあたって知識の検索, 知識埋め込み表現へと変換する機構について詳しく説明する.

知識グラフ前処理　知識グラフ前処理では, 入力単語列 $X = (w_1, ..., w_n)$ に対して追加できる知識を KG に対して検索する. この際, X を分かち書き前の文として扱い, 文に対して mecab-ipadic-NEologd(以降, NEologd)[1] [7] 辞書を使用した MeCab による分かち書きを行う. 分かち書き後の X 中の単語のうち, 名詞のものをキーワードとして KG に存在するか検索を行い, ヒットしたものを追加する知識群 $C = (c_1, ..., c_m)$ とする. 例えば, キーワード群「人間, 山田太郎, SCSK」が KG 上に存在するか検索を行い「人間, SCSK」が存在した場合, m は 2 となる. 追加する知識の順番は入力単語列中での検索キーワードの登場順となる.

知識グラフ埋め込み層　知識グラフ埋め込み層では, 知識グラフ前処理で獲得した知識群 $C = (c_1, ..., c_m)$ を知識埋め込み表現へと変換する. 本研究では, KG 埋め込み手法の 1 つである TransE [8] により, KG 上で登場する知識の埋め込み表現を事前に別途学習し, 獲得している. TransE により獲得した KG の埋め込み行列を用いて, 知識群 $C = (c_1, ..., c_m)$ を知識埋め込み表現へと変換する. 通常, TransE による埋め込み次元数は BERT の単語埋め込み層の埋め込み次元数 d と異なる. したがって, TransE の埋め込み行列による変換後に全結合層を用いて $[a_1, ..., a_m] \in \mathbb{R}^d$ へと変換する. 全結合層は学習可能パラメータとする.

続いて, 入力テキストの単語埋め込み表現と知識埋め込み表現の結合処理の詳細を説明する. 知識埋め込み表現は単語埋め込み表現の後ろに結合する. 位置エンコーディングは結合後の入力系列に対して, BERT と同様に先頭トークンから順に位置情報を付与する. 文のセグメント情報についても同様に, 各文に対し知識群を加えた系列長を 1 文とみなし, 各文が 1 文目か 2 文目であるかの識別情報として付与する.

BERT への入力が 2 文 $X_1 = (w_{11}, ..., w_{1n})$, $X_2 = (w_{21}, ..., w_{2k})$ である場合, X_1, X_2 それぞれに対して KG 検索を行った後, 埋め込み表現 $\{a_{11}, ..., a_{1m}\}$, $\{a_{21}, ..., a_{2l}\}$ へ変換する. 獲得した埋め込み表現を各文の単語埋め込み表現へ結合し, 結合後の各入力系列を特殊トークン [SEP] によって結合した $\{e_{\text{CLS}}, e_{w_{11}}, ..., e_{w_{1n}}, a_{11}, ..., a_{1m}, e_{\text{SEP}}, e_{w_{21}}, ..., e_{w_{2k}}, a_{21}, ..., a_{2l}, e_{\text{SEP}}\}$ に位置エンコーディングとセグメント情報の埋め込み表現を加算したものを BERT への入力とする (n, k は入力単語数, m, l は追加する知識数). ここで, BERT 単語埋め込み層で埋め込み表現へ変換する単語列は, Hugging Face が提供している Transformers [9] の BertJapaneseTokenizer を使用して分かち書きしたものである.

表 1: JGLUE の構成

タスク	データセット	学習用	開発用	テスト用
文章分類	MARC-ja	181,864	5,660	5,654
文ペア分類	JSTS	10,862	1,589	1,457
	JNLI	20,073	1,216	1,218
QA	JSQuAD	57,777	5,082	4,442
	JCommonsenseQA	7,821	1,118	1,119

3 実験

3.1 実験設定

本節では, 提案手法を JGLUE [6] と子育て FAQ データ (CC-BY 4.0 子育てオープンデータ協議会)[2] の両データセットで評価を行い, 提案手法の有効性を検証する. 前処理を含めた実験の詳細は付録 A を参照されたい.

知識グラフ　本実験では, JGLUE, 子育て FAQ データともに, 知識グラフに Wikidata[3] を使用した. 提案手法の知識グラフ埋め込み層で使用する KG の埋め込み行列は, OpenKE [10] で学習済みのものを使用した.[4]

JGLUE　本データは, 一般的な日本語理解能力を測ることを目的とした言語理解ベンチマークである. 2022 年 8 月時点において, 文章分類データセッ

1) 事物 (インスタンス) の関係をグラフで整理した KG に対して, インスタンス名を幅広く検索できるように固有表現に強い NEologd を採用した.

2) https://linecorp.com/ja/csr/newslist/ja/2020/260
3) https://dumps.wikimedia.org/wikidatawiki/entities/
4) http://139.129.163.161/index/toolkits#pretrained-embeddings

表 2: 実験結果

Model	MARC-ja acc	JSTS Pearson/Spearman	JNLI acc	JSQuAD EM/F1	JCommonsenseQA acc	子育て FAQ acc(平均)
BERT	95.31	87.06 / 81.91	84.56	82.67 / 90.75	75.16	91.66
提案手法	**95.68**	**87.58 / 82.70**	**85.88**	81.97 / 90.13	**75.72**	**92.91**

ト (MARC-ja), 文ペア分類データセット (JSTS, JNLI), QA データセット (JSQuAD, JCommonsenseQA) の計 5 つのデータセットが利用可能であり, 本実験ではこれら 5 つのデータで評価を行う. 各データセットは同時点でテストデータが公開されていないため, 学習用/開発用から一部データを抽出しテストデータとして用いた. 各データセットの内訳を表 1 に示す.

子育て FAQ データ　本データは,「自治体における AI チャットボットの利活用促進」を目的に LINE が参画する子育てオープンデータ協議会より公開されている. 本データは, 子育てに関する質問サンプルとその回答サンプル, サンプルの分類カテゴリで構成されており, 本実験では 481 件を使用した. データ数が少ないため, 質問文をカテゴリに分類するタスクの性能を, 5 分割交差検定で評価した.

3.2　実験結果

JGLUE および子育て FAQ データの実験結果を表 2 に示す. JGLUE では, 提案手法が JSQuAD 以外の 4 つのタスクでベースラインを上回る結果となった. 特に, JNLI において+1.32 ポイントと全タスクの中で最も高い精度向上となった.

子育て FAQ データでは, ベースラインと比較して+1.25 ポイントの精度向上であり, ビジネスで利用される実データの文章分類においても提案手法が有効であることを確認した.

3.3　KG の影響確認

子育て FAQ データでの実験について, KG を活用したことによる影響を LIME [11][5] を用いて確認した. LIME とは, 機械学習モデルへ入力する属性群のうちどの属性が予測に寄与したかを可視化分析するツールである. 本節では, 知識追加による影響度の確認を目的とするため, 知識グラフ前処理と同様に MeCab による分かち書き後の単語列を属性群とした. LIME は, 各ラベルに対する予測確率の内訳と, 最も確率が高いラベルとそれ以外のラベルに関して, それぞれの予測結果の要因となった属性を示すグラフを表示する. 図 2a〜3b について, 各図の左側のグラフが各ラベルに対しての予測確率, 右側が要因となった属性のグラフとなっている.

図 2a, 2b は, BERT の予測は不正解で, 提案手法において正解したサンプルの LIME の分析結果である. 入力文が「保育 の しおり が ほしい です .」で分類カテゴリが保育であるサンプルにおいて, BERT では妊娠・出産ラベルに誤分類しているのに対して, 提案手法では正しく予測できている. このとき, 提案手法では, KG から「保育」が知識として追加されている.

続いて, 図 3a, 3b で, BERT の予測では正解して, 提案手法では不正解であったサンプルの LIME の分析結果を確認する. 入力文が「アート スクール は あり ます か ?」で分類カテゴリが施設であるサンプルにおいて, BERT では施設ラベルを正しく予測できているのに対して, 提案手法では保育ラベルを誤って予測している. 図 3a の BERT の分析結果より, 施設ラベルを予測する際に「アート」,「スクール」といった属性を重要視している. 一方, 図 3b の提案手法では, 保育ラベルを予測する際に「スクール」を重要視しており, 施設の語義で使用されていた「スクール」を他語義で解釈していると考えられる. 提案手法では, KG 上でロックバンドを指す「ART-SCHOOL」が知識として追加されており, 追加された知識により入力文中の語義とは異なる解釈がされた可能性がある.

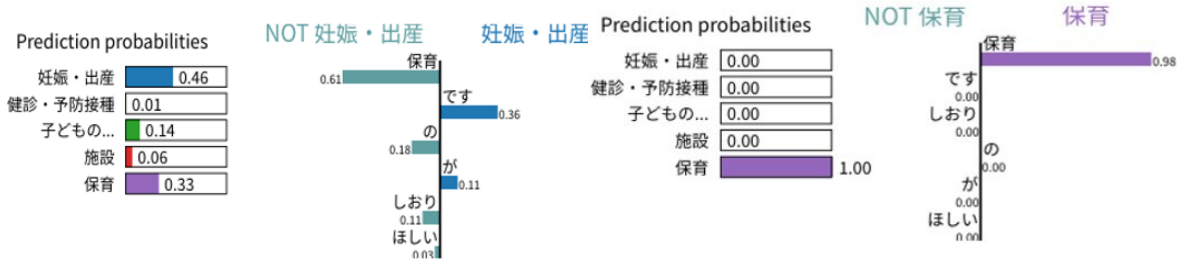
今回の分析により, KG の知識を追加する際に, 正解ラベルと関連度の高い入力文中の単語に対して, 文脈中の語義に沿った知識を追加することで正解率が向上する可能性が示唆された. このことを, 他のデータやタスクも含めてより詳細に分析することは今後の課題である.

4　関連研究

BERT の事前学習またはファインチューニング時に KG の知識を活用し, BERT の言語理解性能を向上させる研究が盛んに行われている.

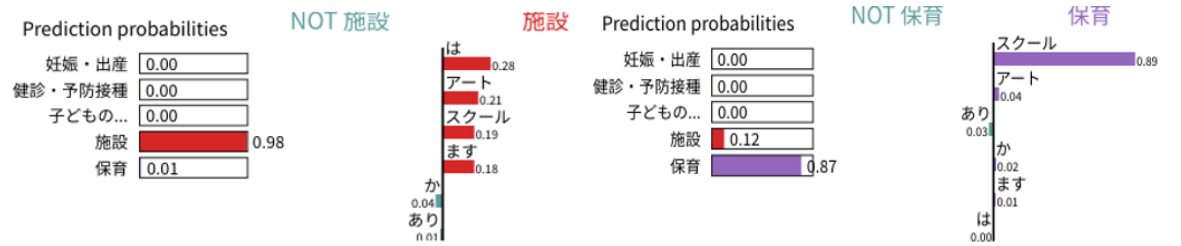
ERNIE [3] は, モデルの事前学習時に入力文に関連

5)　https://github.com/marcotcr/lime



(a) BERT **(b)** 提案手法

図 2: 提案手法で新規に正解した例



(a) BERT **(b)** 提案手法

図 3: 提案手法で新規に不正解になった例

する KG の知識を埋め込むための情報合成層と専用の事前学習タスクを提案している. BERT への入力は, BERT の単語埋め込み層で変換した入力文の埋め込み表現と入力文中の単語に関連する KG 上の知識を TransE で変換した知識埋め込み表現である. このとき, 単語埋め込み表現と知識埋め込み表現はそれぞれ異なるマルチヘッドアテンションに入力されることに留意されたい. ERNIE は, KG 上の知識を活用することでドメイン固有タスクで高い性能を達成することができた. しかし, 知識を追加する際に, 単語の語義曖昧性や同音意義語の問題を解消しておらず, 入力文中の単語に対して適切な知識を追加できないケースが存在する.

KI-BERT [4] は, モデルのファインチューニング時に, KG の知識を考慮した学習を可能とする手法である. 入力文中の単語に対して, 概念情報と語義曖昧性を解消した知識の追加を可能としている. 入力単語に対して概念情報と語義曖昧性を解消した語義情報のどちらを追加するかの判定を行い, 概念情報には ConceptNet [12] の知識, 語義が曖昧な単語には WordNet [13] の知識を追加する. このとき, 追加する知識は, 各 KG について事前に TransE [8] または ConvE [14] で獲得した埋め込み行列を用いて変換され, 提案手法同様に各入力文の直後に追加される. KI-BERT は GLUE [5] の各タスクにおいて ERNIE などの既存の KG を考慮したモデルと比較して大きな精度向上を達成した.

英語データを対象とした BERT モデルにおいて外部知識を活用することで精度改善が達成できるように, 日本語モデルにおいても外部知識を活用することで精度改善が達成できることが報告されている [15, 16]. LUKE [15] は, 文脈情報に基づく新たな単語とエンティティの埋め込み表現の学習方法を提案し, 下流タスクで大きな精度向上を達成した. 笹沢らの研究 [16] では, 製品に対するレビューの分類を行う感情分類モデルにおいて, レビュー文に対してユーザ ID と製品 ID の 2 種類のカテゴリ属性情報をファインチューニング時に追加することで精度が向上することが報告されている.

5 おわりに

本研究では, 日本語の BERT モデルに KG の知識を活用することによる下流タスクの性能向上の可能性を検討した. 実験の結果, 先行研究の英語データを対象とした実験同様に, 日本語においても KG の知識を活用することで, 多くのタスクで精度向上が確認できた. また, 文章分類での BERT と提案手法の LIME を使用した可視化分析により, 入力文中の語義に沿った KG の知識が追加されることが, 性能向上のために重要である可能性が示唆された.

今後は, 先行研究と同様に日本語においても, 入力文中での各単語の語義に適した知識を追加するため, エンティティリンキング手法などを適用することで, 精度が改善されるか確認を行いたい.

参考文献

[1] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 2901–2908, 2020.

[3] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In **Proceedings of ACL 2019**, 2019.

[4] Keyur Faldu, Amit Sheth, Prashant Kikani, and Hemang Akbari. Ki-bert: Infusing knowledge context for better language and domain understanding. **arXiv preprint arXiv:2104.08145**, 2021.

[5] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[6] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会, 2022.

[7] 奥村学佐藤敏紀. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. NLP2017–B6–1. 言語処理学会, 2017.

[8] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. **Advances in neural information processing systems**, Vol. 26, , 2013.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, Online, October 2020. Association for Computational Linguistics.

[10] Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. Openke: An open toolkit for knowledge embedding. In **Proceedings of EMNLP**, 2018.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016**, pp. 1135–1144, 2016.

[12] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In **Thirty-first AAAI conference on artificial intelligence**, 2017.

[13] George A Miller. Wordnet: a lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[14] Tim Dettmers, Minervini Pasquale, Stenetorp Pontus, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In **Proceedings of the 32th AAAI Conference on Artificial Intelligence**, pp. 1811–1818, February 2018.

[15] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics, 2020.

[16] 笹沢裕一, 岡崎直観. 属性情報を追加した事前学習済みモデルのファインチューニング. 言語処理学会第 27 回年次大会, 2021.

A　実験の詳細説明

JGLUE　各データセットの詳細を以下に示す.

MARC-ja は, amazon における商品レビューの positive, negative の 2 値分類タスクである.

JSTS は, 2 文の意味的な類似度を推定するタスクである. 正解の類似度は 0(意味が完全に異なる)〜5(意味が等価) の 6 段階となる.

JNLI は, 前提文と仮設文のペアに対して推論関係を予測するタスクである. 推論関係は, 含意, 矛盾, 中立の 3 つである.

JSQuAD は機械読解の QA データセットの 1 つである. 与えられた文章を読み, 内容に関する質問とその答えを文章中から抽出するタスクとなっている.

JCommonsenseQA は, 常識推論能力を評価するための 5 択式の QA データセットである.

本実験では, 開発用データに対して最も性能の良いモデルに対してテスト用データによる評価を行った. 実験に当たって, 公開されているデータから一部データを抽出し, 開発用 (またはテスト用) データとした. データ分割後の各データセットの内訳を表 1 に示す. JNLI は開発用データから一部データを抽出し, 抽出したデータをテスト用とした. JNLI 以外のデータについては, 学習用データから一部データを抽出し, 抽出したデータを開発用とした. また, 開発用データとして公開されているものを本実験ではテスト用データとして使用した.

子育て FAQ データ　本データは, サンプル問合せ文とサンプル応答文, 各サンプルのカテゴリの組み合わせで用意されており, 各自治体に合わせた情報を入れることが可能となっている. 本実験では, 広島県に関する情報を入れたものを使用した. 各カテゴリごとのサンプル数について最大 145 件, 最小 1 件と大きな偏りがあるため, 学習時のノイズ除去の観点から, サンプル数が最低 60 件以上のカテゴリのものを使用した. 前処理後のカテゴリ数は 5 となった. また, 前処理後のデータ件数が全体で 481 件と少量であるため, 本実験では, 5 分割交差検定で実験を行い, JGLUE での実験と同様に開発用データに対して最も精度が高いモデルに対してテスト用データで評価を行った.