エンティティの言語間リンクに基づく多言語モデルの構築

西川 荘介[1,2]　山田 育矢[2,4]　鶴岡 慶雅[1]　越前 功[1,3]
[1] 東京大学 大学院情報理工学系研究科　[2] Studio Ousia
[3] 国立情報学研究所　[4] 理化学研究所
{sosuke-nishikawa,iechizen}@nii.ac.jp
ikuya@ousia.jp　tsuruoka@logos.t.u-tokyo.ac.jp

1　はじめに

様々な言語処理タスクにおいて、資源の豊富な言語（原言語）における教師データで学習し、他の言語（目的言語）における推論で性能を向上させる言語間転移学習が研究されている. 近年これらの手法の多くは事前に多言語の大量のコーパスで学習された事前学習済み汎用多言語モデルを用いて実現される [1, 2, 3]. しかし、言語体系が異なる言語対や事前学習時に用いる目的言語のコーパスの量が比較的少ない場合、汎用多言語モデルを用いた言語間転移学習で十分な性能を発揮できない場合があることが報告されている [4].

本研究では、知識ベース (Wikipedia) に紐づけられたエンティティ表現を用いて汎用多言語モデルの性能向上を試みる. エンティティは曖昧性を持たず文書に関連する情報を持つという特徴があり、エンティティを用いて文書分類タスクやエンティティ型推定タスクを解く手法が提案されている [5, 6, 7].

本研究では、エンティティ間の言語間リンクを利用することで目的言語の推論データから抽出されたエンティティ群を原言語のエンティティ群に変換することで言語間の転移学習を実現する（図 1）. さらに、汎用多言語モデルから得られる文脈表現から学習される注意機構によりエンティティ表現の重みつき和を計算することで、該当文書に関連するエンティティ表現を優先的に考慮するような機構を導入する. このエンティティ表現を汎用多言語モデルから得られる表現に付加的に加算することで性能の向上を試みる. この手法では原言語で学習したエンティティ表現を推論に直接用いることで言語間転移学習を可能とする. また、既存の多言語モデルに容易に組み込むことができ、追加の学習のデータを必要とせず言語対ごとに再学習する必要がない.

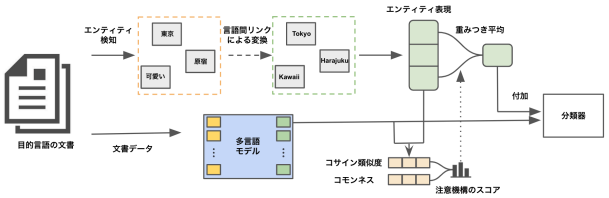

図 1　提案手法の概要図. 図は目的言語における推論時の処理を表す. 学習時は言語間リンクによる変換を行わない.

実験では、文書分類タスクとエンティティの型推定タスクにて、原言語を英語とし目的言語を 7 言語および 10 言語で評価を行い、ベースラインとなる汎用多言語モデルと比較して性能の向上を確認した.

2　関連研究

2.1　言語間転移学習

言語間転移学習を実現するにあたり、機械翻訳を用いた手法や多言語分散表現を用いた手法が提案されている [8, 9]. 近年では Multilingual-BERT (M-BERT) [1] 等に代表される多言語の大規模言語コーパスを用いて事前訓練された汎用多言語モデルが提案され [1, 2, 3]、このモデルに基づく多言語モデルが様々な言語間転移タスクで高い精度を達成している. しかし、これらの手法では異なる言語体系を持つ言語対や事前学習用コーパスにあまり現れない言語に対して、性能が十分に発揮されないことが報告されている [4].

この問題に対し、目的言語の追加データを用いて汎用多言語モデルの性能の向上を試みる手法が提案されている. 目的言語のラベルなしコーパスを利用することで言語とドメインの差を埋める学習や [10]、原言語で学習したモデルを用いて目的言語のラベルなしコーパスに擬似ラベルを付与しデータ

拡張する手法 [11]、対訳コーパスを用いて再学習する手法 [3] 等がある. しかしこれらの手法では言語対ごとに追加データを利用して多言語モデルを再学習する必要があり、学習コストがかかる. また、再学習されたモデルは該当する 1 つの目的言語で性能を発揮するが、複数の言語に対して同時に性能を発揮する多言語モデルの利便性が失われている.

提案手法では原言語で学習したエンティティ表現を直接目的言語での推論に用いるため、モデルは言語対に依存せず性能を発揮することが期待される. また、追加データを必要とせず単一モデルの学習のみでの言語間転移の性能向上が可能となる.

2.2 エンティティ表現

エンティティ表現とは Wikipedia 等の知識ベースに紐づくエンティティを表すベクトル表現であり、多くの自然言語処理タスクに応用されている [5, 6, 7, 12]. Wikipedia2Vec [13] はエンティティ表現を学習する手法の 1 つであり、skip-gram モデル [14] に基づき Wikipedia ページにおけるエンティティの周辺単語の予測と Wikipedia のページ間リンクでつながっているエンティティの予測を行うことでエンティティ表現を学習する. 本研究ではこの Wikipedia2Vec を用いてエンティティ表現を事前に学習する.

エンティティ表現の応用例の一つとして文書分類タスクがあり、文書中から文書に関連するエンティティ群を優先するような注意機構により重み付けをしたエンティティ表現の和を用いて、単言語の文書分類を解く手法が提案されている [5]. この研究では注意機構を学習する際、単語分散表現の平均ベクトルとエンティティ表現のコサイン類似度を素性として用いているが、本研究では多言語事前学習モデルの隠れ表現とエンティティ表現のコサイン類似度を素性として注意機構を学習する.

また、エンティティ表現を汎用言語モデル BERT [1] に組み込む関連研究として、エンティティ表現用の追加のエンコーダーを導入し再度事前学習を行う手法 [15, 16] や文書中のエンティティ名の近傍に知識ベースに紐づくエンティティを挿入する手法 [17]、エンティティ表現を事前学習により直接獲得する手法 [18, 19] 等が提案されている. 本研究では多言語タスクの性能向上を目的としており、ある多言語タスクの学習と同時にエンティティ表現をモデルに導入する機構を学習するため、追加の事前学習を必要としない. また、多言語モデルの内部構造を改変することなく外部から付加的にエンティティ表現を導入するため、複数提案されている既存の多言語モデルに対して本手法を適用することができる.

3 提案手法

本節ではエンティティ表現を汎用多言語モデル M-BERT に組み込む手法を提案する.

エンティティの処理

まず文書から文書中に現れるエンティティ群を抽出する. 本研究では知識ベースとして Wikipedia を用い、Wikipedia におけるページ内リンク情報 [20] からエンティティ名（"東大"）とエンティティ（東京大学）を紐づけるエンティティ辞書を構築する. このエンティティ辞書を参照して文書中のすべてのエンティティ名から対応するエンティティ候補を全て抽出する[1)]. また、あるエンティティ名がページ内リンクである確率（リンク確率）とあるエンティティ名が特定のエンティティを示す確率（コモンネス）も事前に獲得する.

Wikipedia では同じエンティティに関するページが複数の言語で作成されており、同一のエンティティは言語をまたいで言語間リンクで接続されている. 目的言語における推論時は、文書から抽出される目的言語のエンティティ群をエンティティの言語間リンクを用いて原言語のエンティティ群に変換する.

注意機構の学習

次に文書中から抽出されたエンティティ群それぞれに、対応する事前学習済みエンティティ表現を割り当てる. 本手法ではエンティティ辞書を参照してエンティティ名に紐づく全てのエンティティを抽出しているため、該当文書には関連しない知識ベースに紐づけられたエンティティも抽出している場合がある. この問題を解決するため、文書に関連するエンティティ表現を優先するような注意機構を導入する. 具体的には以下のようにエンティティ表現 $\boldsymbol{v_e} \in \mathbb{R}^d$ の重みつき和を計算することで文書に関連したエンティティに基づく特徴量 $\boldsymbol{z_{entity}} \in \mathbb{R}^d$ を

1) 本研究では文書からエンティティ群を抽出する際にエンティティの曖昧性解消の処理は行わない.

表 1 エンティティの型推定タスクに用いたデータセット

言語	例数
English (en)	439,354
French (fr)	318,828
German (de)	274,732
Japanese (ja)	920,443
Chinese (zh)	267,107
Italian (it)	270,295
Russian (ru)	253,012
Spanish (es)	257,835
Hindi (hi)	30,547
Thai (th)	59,791
Arabic (ar)	73,054

表 2 文書分類タスクに用いたデータセット

言語	訓練	検証	テスト
English (en)	1000	1000	-
fr, de, ja, zh, it, ru, es	-	-	4000

得る.

$$\boldsymbol{z_{entity}} = \sum_{i=1}^{K} a_{e_i} \boldsymbol{v_{e_i}} \tag{1}$$

ここで $a_e \in \mathbb{R}$ はエンティティ e に対応する注意機構の重みであり、以下の式によって $\boldsymbol{a} = [a_{e_1}, a_{e_2}, \cdots, a_{e_K}]$ を学習する.

$$\boldsymbol{a} = Softmax(\boldsymbol{W_a}^\top \boldsymbol{\phi}) \tag{2}$$

$$\phi(e, D) = cosine(\boldsymbol{h}, \boldsymbol{v_e}) \oplus p \tag{3}$$

ここで $\boldsymbol{W_a} \in \mathbb{R}^2$ は重みベクトルである. $\boldsymbol{\phi} = [\phi(e_1, D), \phi(e_2, D), \cdots, \phi(e_K, D)]$ は文書 D から抽出されたエンティティ群と D の関連度合いを測る素性であり、M-BERT の入力列の [CLS] トークンに対応する最終の隠れ表現 $\boldsymbol{h} \in \mathbb{R}^d$ とエンティティ表現 $\boldsymbol{v_e}$ のコサイン類似度にコモンネス p を連結させたベクトルである.

汎用多言語モデルへの付加

BERT に基づく汎用言語モデルで文書分類タスクを解く際はこの隠れ表現 $\boldsymbol{h}$ を付加的に接続された分類器に入力し、クラス c の確率を推論する. 提案手法ではこの隠れ表現 $\boldsymbol{h}$ に、$\boldsymbol{z_{entity}}$ を足し合わせた表現を分類器に入力しモデルを学習する.

$$p(c \mid \boldsymbol{h}, \boldsymbol{z_{entity}}) = Classifier(\boldsymbol{h} + \boldsymbol{z_{entity}}) \tag{4}$$

4 実験

本研究では文書分類タスクとエンティティの型推定タスクにて提案モデルの評価を行った. 両タスクでは英語を訓練データ、検証データとし、その他の言語をテストデータとした.

本手法は任意の多言語モデルに適用可能であるが、実験では M-BERT の既存実装[2] をベースラインとして用いて提案手法の有効性を検証した.

4.1 実験設定

文書分類タスク

文書分類タスクとは文章をトピックに対して分類するタスクであり、データセットとして MLDoc [21] を用いた. この MLDoc は文書を CCAT (Corporate/Industrial)、ECAT (Economics)、GCAT (Government/Social)、MCAT (Markets) のいずれかのカテゴリに分類するマルチクラス分類タスクであり、8 言語での文書データが収録されている。データセットの詳細を表 2 に示す.

実験では M-BERT モデルに対して全結合層と Softmax 層を接続し Cross-entropy 損失を最適化することで学習する. また学習は異なる乱数シードで 10 回行い、その正解率の平均を示す.

エンティティの型推定タスク

エンティティの型推定とは Wikipedia のページをカテゴリに分類するタスクであり、データセットとして SHINRA2020-ML [3] を用いた. このタスクは Wikipedia のページに対して約 220 種類の拡張固有表現[4] のラベルを付与するマルチラベル分類タスクである. 公開されているデータセットの中で MLDoc と同じ 8 言語に加えて、英語と言語体系が大きく異なる言語群（hi、ar）と M-BERT の事前学習データに含まれていない言語（th）を選択した. データセットの詳細を表 1 に示す. 検証データとして学習データの 5% をラベルの比率を維持して抽出したものを利用した.

実験では M-BERT モデルに対して全結合層と Sigmoid 層を接続し Binary-cross-entropy 損失を最適化することで学習した. 推論時は Sigmoid 層の出力が 0.5 以上になるラベルを予測ラベルに追加した.

2） https://github.com/huggingface/transformers
3） http://shinra-project.info/shinra2020ml/
4） Extended Named Entity homepage: https://ene-project.info

表 3　文書分類タスクにおける正解率

モデル	en	fr	de	ja	zh	it	ru	es	ave.
M-BERT	94.1	79.8	74.6	70.2	72.7	67.2	66.4	75.7	75.1
M-BERT + BoE (-emb)	94.0	80.9	76.1	71.3	**76.5**	68.0	67.5	76.2	76.3
M-BERT + BoE (-att)	93.5	77.4	72.2	68.4	71.8	66.0	64.6	74.6	73.6
M-BERT + BoE	94.2	**83.6**	**79.2**	**71.5**	74.4	**71.0**	**68.4**	**77.4**	**77.5**

表 4　エンティティの型推定タスクにおける F1 値のマイクロ平均

モデル	en (dev)	fr	de	ja	zh	it	ru	es	hi	th	ar	ave.
M-BERT	89.1	76.5	84.8	85.8	83.5	86.5	83.5	86.6	60.6	61.7	62.8	75.7
M-BERT + BoE (-emb)	89.0	**80.2**	84.9	86.0	83.7	87.0	83.7	86.7	63.5	63.0	62.0	78.1
M-BERT + BoE (-att)	78.5	48.3	61.2	62.8	57.7	63.2	56.9	61.4	38.6	44.3	47.2	54.2
M-BERT + BoE	89.1	73.6	**85.3**	**86.2**	**84.2**	**87.5**	**84.7**	**86.8**	**66.5**	**63.7**	**67.8**	**78.6**

また、SHINRA2020-ML での評価指標 [22] に倣い、F1 値のマイクロ平均を評価指標として用いた.

エンティティの処理

2019 年 1 月版の英語 Wikipedia コーパスから Wikipedia2Vec [13] を用いてエンティティ辞書の作成とエンティティ群の抽出を行った. この際、リンク確率が 0.01 以上かつコモンネスが 0.05 以上のエンティティを抽出した. エンティティの言語間リンクは Wikidata [5] から取得した.

また、提案手法ではエンティティ表現 $\boldsymbol{v_e}$ を事前に学習した表現で初期化した. エンティティ表現の学習には Wikipedia2Vec を用い、同じ版の英語 Wikipedia コーパスからベクトルの次元数を 768、その他のパラメータはデフォルト値に設定して学習した.

その他の設定

全ての実験において学習率は 2×10^{-5} に設定し、パラメータ更新の最適化アルゴリズムは AdamW [23] を用いた. バッチサイズは文書分類タスクでは 32、エンティティの型推定タスクでは 256 に設定して学習した. また、検証用データにおいてそれぞれの評価指標における性能が改善されなくなるまで学習を行った.

また、エンティティ表現を事前学習する有効性を検証するため、エンティティ表現をランダムに初期化したモデル（M-BERT + BoE (-emb)）を学習した. さらに、注意機構の有効性を検証するため、注意機構を取り除いたモデル（M-BERT + BoE (-att)）を学習した.

5） https://dumps.wikimedia.org/wikidatawiki/entities/

4.2　結果

表 3, 4 に文書分類タスクとエンティティの型推定タスクの実験結果を示す. どちらのタスクにおいても多くの目的言語において提案手法がベースラインを上回る性能を示した. 特にエンティティ型推定タスクにおける英語と言語体系的に大きく異なる言語群（hi、ar）において 5 ~ 6 ポイントの性能の向上を確認した.

また、多くの言語対で事前学習済みエンティティ表現による初期化を行うモデルの方が性能が高いが、一部の言語対では初期化を行わない方が性能が高くなる場合があることが分かった. さらに、注意機構を利用しなかったモデルはベースラインよりも悪い性能となっており、注意機構による重み付けが有効に働いていることを確認した.

5　おわりに

本研究ではエンティティの言語間リンクにより原言語に変換したエンティティ表現を付加的に用いることで、汎用多言語モデルにおける言語間転移学習の性能を向上させる手法を提案した.

実験では言語間の文書分類タスク、エンティティの型推定タスクにおいてベースラインである M-BERT と比較し提案手法の優位性を示した. また、注意機構や事前学習済みエンティティ表現による初期化が有効に機能していることを確認した.

本研究ではベースラインとして M-BERT を用いているが、他の事前学習済み汎用多言語モデルや多言語分散表現を用いたモデルにおいても提案手法が有効に機能するか検証することが今後の展望として考えられる.

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)*, pages 4171–4186, June 2019.

[2]Sebastian Ruder, Anders Søgaard, and Ivan Vulić. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, July 2019.

[3]Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7059–7069, 2019.

[4]Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, July 2019.

[5]Ikuya Yamada and Hiroyuki Shindo. Neural attentive bag-of-entities model for text classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 563–573, November 2019.

[6]Ikuya Yamada, Hiroyuki Shindo, and Yoshiyasu Takefuji. Representation learning of entities and documents from knowledge base descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 190–201, August 2018.

[7]Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2915–2921, 2017.

[8]Kui Xu and Xiaojun Wan. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, September 2017.

[9]Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904, April 2017.

[10]Guokun Lai, Barlas Oguz, Yiming Yang, and Veselin Stoyanov. Bridging the domain gap in cross-lingual document classification. *arXiv preprint arXiv:1909.07009*, 2019.

[11]Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, November 2019.

[12]Nina Pörner, Ulli Waltinger, and Hinrich Schütze. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. *arXiv preprint arXiv:1911.03681*, 2019.

[13]Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30, October 2020.

[14]Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[15]Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, July 2019.

[16]Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, November 2019.

[17]Nina Poerner, Ulli Waltinger, and Hinrich Schütze. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, November 2020.

[18]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)*, pages 6442–6454, November 2020.

[19]Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, November 2020.

[20]Stephen Guo, Ming-Wei Chang, and Emre Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030, June 2013.

[21]Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018.

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

[23]Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

A　付録

謝辞

本研究は JSPS 科研費 JP16H06302, JP18H04120 および JST CREST JPMJCR18A6, JPMJCR20D の助成を受けたものです.