テキストと非テキストデータからの同時事前学習

友利 涼 † 亀甲 博貴 †† 森 信介 ††

† 京都大学情報学研究科 †† 京都大学学術情報メディアセンター
{tomori.suzushi.72e@st, kameko@i, forest@i}.kyoto-u.ac.jp

1 はじめに

大量の生テキストからモデルを事前学習する手法がいくつか提案されており [2]、ファインチューニングすることで様々なタスクで高精度を達成している。生テキストから事前学習されたモデルは、ある種の実世界知識を保有していることが報告されている。文献 [4] では、統語情報などの言語知識を保有していることが示されており、文献 [13] では、実世界に接地された知識を保有していることを質問応答データセットや知識ベースを用いて示されている。

従来の事前学習手法ではテキストのみからモデルを学習していたが、実世界知識にはテキストのみでは表現しきれないものも多い (視覚的な情報や触覚的な情報など)。多くの自然言語処理タスクは本質的にテキストのみでは表現しきれない様々な実世界知識を必要とするが、従来の自然言語処理システムはこの問題を考慮できてないことが多い。そのため、様々な種類のデータから幅広い実世界知識を事前学習により獲得する手法が求められている。

本論文では、幅広い実世界知識を獲得するために、テキストと非テキストデータからそれらの関係性を事前学習する手法を提案する。それらの関係性を学習したモデルは、様々なテキスト解析タスクに有用だと考える。テキストと非テキストデータの両方を扱うため、テキストエンコーダと非テキストデータエンコーダを統合するモデルを提案する。

本論文では、2 つのドメイン・6 つのテキスト解析タスクで実験・評価を行う。また、その他の貢献として、応用タスクに合わせた解析モデルも提案する。実験結果より、提案手法はテキストのみから事前学習する手法よりも高精度を達成したことを示す。さらに、従来手法と比較し提案手法は、実世界に接地された知識を保有していることを知識ベースを用いて示す。

2 事前学習

本論文では、テキストと非テキストデータから事前学習する手法を提案する。図 1 にその概要を示す。入力文 $\boldsymbol{x} = ([\mathrm{CLS}], x_1, x_2, ..., x_N, [\mathrm{SEP}])$ とその文に対応する非テキストデータが与えられた時、モデルは分散表現列 $\boldsymbol{h} = (\boldsymbol{h}_{[\mathrm{CLS}]}, \boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N, \boldsymbol{h}_{[\mathrm{SEP}]})$ を出力する。ここで、[CLS] と [SEP] はそれぞれ文頭と文末を示す特別な記号であり、 $\boldsymbol{h}_n$ は x_n の分散表現である。

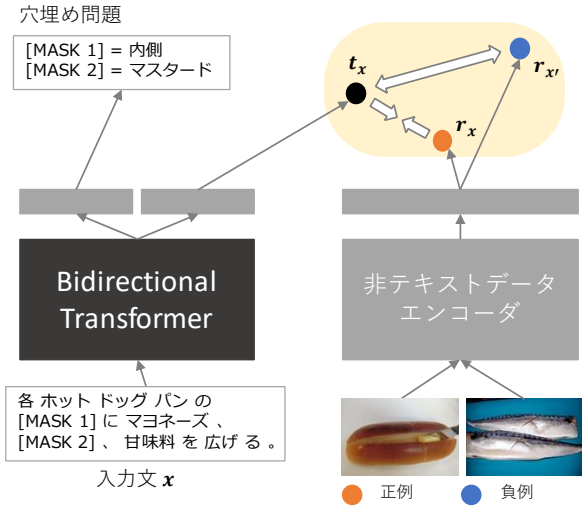

図 1: 提案手法の概要

テキストのエンコードには BERT [2] を用いる。以下では、テキストからの事前学習手法について簡単に説明する。まず、入力文から一様に 15%の単語選ぶ。選ばれた単語の内 80%を特別な記号 [MASK] に、10%をランダムな単語に置き換え、残りの 10%は特に変更はしない。変更した文をモデルに入力し、得られた分散表現から元の単語を予測することで学習を行う。この時の目的関数 L_{MLM} はクロスエントロピー誤差の対数とする。BERT の元論文では、さらに次文予測機構も用いて事前学習している。しかし、提案手法では次文予測機構を採用しない。

2.1 テキストと非テキストデータからの同時事前学習

上記のテキストからの事前学習と、テキストとそれに対応した非テキストデータの関係性の学習を同時に行うために、 BERT と非テキストデータエンコーダを統合したモデルを提案する。テキストの分散表現と非テキストデータの分散表現を同一の潜在空間に写像し、それらの距離を深層距離学習手法を用いて最小化することで学習を行う。

文頭の特別な記号の分散表現 $\boldsymbol{h}_{[\mathrm{CLS}]}$ を全結合層に入力し、文の分散表現 $\boldsymbol{t}_{\boldsymbol{x}} \in \mathbb{R}^k$ を得る。非テキストデータを非テキストデータエンコーダに入力する。エンコーダから得られたベクトルを全結合層に入力し、

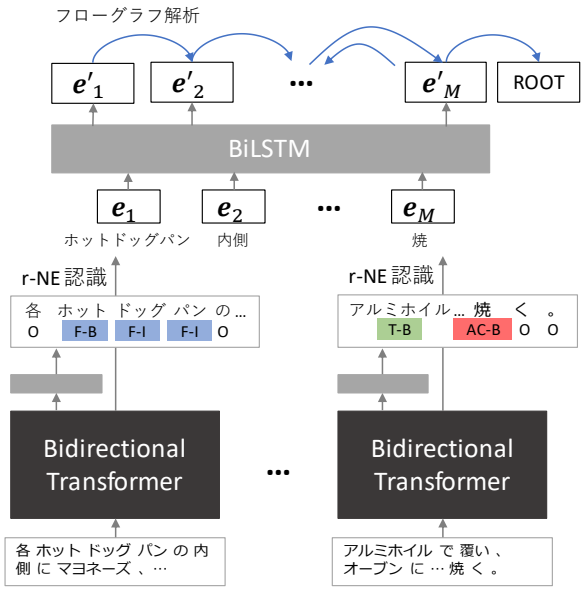

図 2: フローグラフ解析モデル

分散表現 $\boldsymbol{r_x} \in \mathbb{R}^k$ を得る。この時、エンコーダは適切なデータセットやタスクで事前学習したものを用いることができる (非テキストデータが画像の場合、画像分類タスクなどで事前学習可能)。

目的関数 L_{R} には、深層距離学習によく用いられる improved triplet loss [1] を用いる。

$$L_{\mathrm{R}} = \max\left(0, d(\boldsymbol{t_x}, \boldsymbol{r_x}) - d(\boldsymbol{t_x}, \boldsymbol{r_{x'}}) + \alpha\right) + \max\left(0, d(\boldsymbol{t_x}, \boldsymbol{r_x}) - \beta\right)$$

ここで、d は距離を計算する関数であり、本論文では L_2 norm を用いる。α と β はハイパーパラメータであり、正の値を設定する。$\boldsymbol{r_{x'}}$ は入力文 $\boldsymbol{x}$ に対応してない非テキストデータの分散表現であり、セミハードネガティブサンプリング手法により選ばれる。まず、一様に非テキストデータを複数個サンプリングし、その中から $d(\boldsymbol{t_x}, \boldsymbol{r_x}) \leq d(\boldsymbol{t_x}, \boldsymbol{r_{x'}}) < d(\boldsymbol{t_x}, \boldsymbol{r_x}) + \alpha$ の条件を満たす非テキストデータの分散表現を 1 つ選ぶ。深層距離学習手法を用いた理由は、関係性を学習するための適切な手法であるためと、ファインチューニングせずとも潜在空間上でテキスト・非テキストデータの検索が行えるため[1]、様々なタスクに応用可能なためである。最終的な目的関数は $L = L_{\mathrm{MLM}} + L_{\mathrm{R}}$ とする。

3 ファインチューニング

本論文では、レシピドメインの 2 つのテキスト解析タスク、ゲーム解説ドメインの 4 つのテキスト解析タスクを用いて提案手法の有効性を調査する。本節では、それぞれのタスクとその解析モデルについて説明する。各解析モデルは事前学習されたテキストエンコーダをモデルの一部に組み込んでいる。

[1]予備実験では、テキストから対応する非テキストデータの検索、非テキストデータからテキストの検索を行い、類似手法である文献 [6] のモデルよりも検索精度が高かった。

3.1 レシピドメイン

レシピコーパス [10] は cookpad から収集された料理レシピから構成されており、 元々のサイトにおいて、一部のレシピには理解補助のための説明画像が付与されている。レシピコーパスを用いた理由は、レシピ理解にはテキストのみでは獲得しきれない様々な実世界知識が必要となるためと、レシピテキストとそれに対応する画像が大量に入手できるためである。

レシピ固有表現 (r-NE) として、食材名や道具名など、8 種類のレシピドメインに固有な表現が定義されている。各単語に `BIO` ラベル形式で固有表現タイプがアノテーションされており、 `B` はある固有表現の最初の単語を、 `I` は同種の固有表現の継続、 `O` は固有表現以外の単語を表す。

フローグラフはラベル付きの無閉路有向グラフであり、手続き文書理解の 1 つの形として定義されている。グラフの各頂点はレシピに出現した r-NE であり、その r-NE 間の関係性がラベル付きの辺で表される。辺のラベルとして、13 種類が定義されている。

図 2 にフローグラフ解析のモデルの概要を示す。r-NE 認識は系列ラベリング問題として扱う。まず、入力文 $\boldsymbol{x}$ をモデルに与え、分散表現列 $\boldsymbol{h}$ を得る。分散表現列を条件付き確率場 (CRF) [5] 層に入力し、 r-NE のラベル系列を出力する。フローグラフ解析モデルは、まず、上記の方法でレシピから r-NE を抽出し、 r-NE の分散表現列を得る。

$$\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, .., \boldsymbol{e}_M), \boldsymbol{e}_m = [\sum_{m_b \leq i < m_e} \boldsymbol{h}_i : \boldsymbol{l}_m]$$

ここで、 m 番目の r-NE は m_b 番目の単語から m_e 番目の単語で構成されており、 $\boldsymbol{l}_m$ はその r-NE のタイプの分散表現である。r-NE の分散表現列を両方向の LSTM (BiLSTM) に入力し、 $\boldsymbol{e'} = (\boldsymbol{e'}_1, \boldsymbol{e'}_2, .., \boldsymbol{e'}_M)$ を得る。その後、 u 番目の r-NE から v 番目の r-NE へと辺が張られている確率を以下の式で計算する。

$$\frac{\exp\left\{F_{rom}(\boldsymbol{e'}_u) \cdot T_o(\boldsymbol{e'}_v)\right\}}{\sum_{v' \in \{M\}-u} \exp\left\{F_{rom}(\boldsymbol{e'}_u) \cdot T_o(\boldsymbol{e'}_{v'})\right\}}$$

ここで、F_{rom} と T_o はそれぞれ同一次元のベクトルを返す全結合層である。辺ラベルの計算時には、フィードフォワードニューラルネットワークを用い、$\boldsymbol{e'}_u$ と $\boldsymbol{e'}_v$ を連結したベクトルを入力することで各ラベルの出現確率を計算する。推論時には、最大全域木のアルゴリズム [9] を用いる。

3.2 ゲーム解説ドメイン

ゲーム解説コーパス [11, 8] は将棋の解説文から構成されており、各解説文には将棋の局面情報が付与されている。ゲーム解説コーパスを用いた理由は、解説文の理解には将棋そのものや局面に関する知識が必要

表 1: コーパス諸元

レシピ	レシピ数	文数	単語数	画像数	r-NE 数	辺数
事前学習	102,576	589,837	13,944,811	441,855	-	-
学習	245	2,065	30,770	-	9,414	10,227
開発	32	285	4,090	-	1,244	1,347
テスト	30	243	3,374	-	1,076	1,160

ゲーム	文書数	文数	単語数	局面数	s-NE 数	モダリティ表現数	事象数	事実性のラベル総数
事前学習	6,505	272,756	11,021,242	722,470	-	-	-	-
学習	2	1,325	23,370	307	6,658	1,014	3,231	1,982
開発	1	247	4,073	142	1,596	196	600	390
テスト	6	469	6,774	616	2,103	413	1,093	719

表 2: レシピ解析の結果 (Micro F-score)

	r-NE	フローグラフ	
		辺のみ	辺ラベル
ベースライン [7]	-	78.8	69.8
BERT (正解 r-NE あり)	-	80.4	73.3
提案手法 (正解 r-NE あり)	-	**81.1**	**74.2**
BERT	89.6	69.8	64.8
提案手法	**90.2**	**73.5**	**68.1**

なためと、非テキストデータとして画像以外の情報である局面情報を扱えるためである。

ゲーム固有表現 (s-NE) として、戦型名や囲い名など、21 種類の将棋解説文に固有な表現が定義されており、`BIO` ラベル形式でアノテーションされている。

モダリティ表現は次に述べる事象クラスや事実性を示唆する表現であり、8 種類の意味ラベルが定義されている。`BIO` ラベル形式でアノテーションされている。

事象クラスとして、事象らしさ関連や事象に対する情報発信者の態度関連の 8 種類のラベルが定義されている。事象クラスのラベルは、事象を構成する核の述語 1 単語のみに付与されており、それ以外の単語には `O` ラベルが付与される。

事実性には、事象の (情報発信者の主観的な) 成否判断とその確信度として、6 種類のラベルが定義されている。事実性のラベルは、事象クラスが `EVe` の事象に対して付与されており、その他については `O` ラベルが付与されている。

上記 4 つのタスクを系列ラベリング問題として扱う。解析モデルはレシピドメインにおける r-NE 認識と同様のモデルを用いる。

4 評価

4.1 実験設定

まず、日本語の Wikipedia データ約 2,000 万文から BERT を事前学習した。この時の単語分割には KyTea [2] を用い、得られた単語をさらにサブワードに分割した [15]。サブワードの語彙サイズは 32,000 に設定した。

レシピドメインにおける事前学習には、レシピと説明画像が大量にある Cookpad Image Dataset [3] を用いた。この時、テキスト側のエンコーダの初期値として、上記の日本語の Wikipedia データから学習した BERT のパラメータを用いた。画像側のエンコーダとして、画像分類タスクで事前学習した PNASNet-5-Large [3] を用いた。また、比較のために、同コーパスのテキストのみを用いて事前学習した BERT を構築した。ハイパーパラメータ等の設定は提案手法と同様のものにした。

ゲーム解説コーパスでは、局面情報エンコーダを 2 種類用意し、別々のモデルを構築した。1 つは stacked-auto-encoder (SAE) を、もう 1 つは AlphaZero [14] と同様の畳み込みニューラルネットワーク (CNN) を用いた。この 2 つのネットワークはプロの将棋の棋譜から事前学習した。それ以外の設定はレシピドメインと同様にした。

実験に用いたコーパスの諸元を表 1 に示す。ファインチューニング時の各ハイパーパラメータは開発データを用いて調整した。これらの実験では、ファインチューニング時やテスト時には非テキストデータを入力として与えていない。本論文では、ファインチューニング時にシードを変更しながら各実験を 10 回行い、その平均を提示している。

4.2 結果と分析

表 2 にレシピ解析の結果を示す。r-NE 認識とフローグラフ解析の 2 つのタスクにおいて、ベースラインであるテキストのみから学習した BERT よりも提案手法の精度が高かった。表 3 にゲーム解説文解析の結果を示す。提案手法はわずかながらベースラインを上回った。提案手法は各タスク・各評価指標で最高精度を示したが、局面情報エンコーダによる大きな違いは見られなかった。

r-NE 認識において、「しんなり」や「しなっと」などの食材の状態名 `Sf` の認識精度がテキストのみから事前学習した手法を大きく上回った。これら表現は基本的にロングテールな分布を持つためテキストのみから

[2] http://www.phontron.com/kytea/index-ja.html

[3] https://github.com/Cadene/pretrained-models.pytorch

表 3: ゲーム解説文解析の結果

	s-NE			モダリティ表現			事象クラス			事実性		
F-score	Micro	Macro	Span	Micro	Macro	Span	Micro	Macro	Span	Micro	Macro	Span
CRF	78.5	64.5	91.6	65.5	45.4	65.3	72.5	41.9	95.1	73.3	40.1	93.5
BiLSTM	81.1	70.0	94.1	67.4	53.0	72.5	76.0	44.8	96.6	78.0	44.4	96.0
BERT	88.5	72.5	95.0	**68.2**	53.5	72.9	78.1	46.8	97.0	78.1	42.9	96.4
提案手法 (SAE)	88.3	74.0	95.1	**68.2**	53.6	**73.2**	**78.4**	**48.4**	96.7	**78.4**	**47.6**	97.5
提案手法 (CNN)	**88.6**	**74.5**	**95.4**	67.4	**54.5**	72.9	78.0	**48.4**	**97.5**	78.0	46.6	**98.4**

表 4: 料理オントロジー [12] による評価 (ラベルごとの F-score と Micro F-score)

	親	子	兄弟	その他	合計
入力ペア数	250	250	250	250	1,000
BERT	77.7	78.4	63.8	61.3	69.2
提案手法	**79.8**	**82.5**	**64.8**	**63.3**	**71.3**

の学習では認識することが難しいが、画像を間接的に見ることで認識精度が向上したと考えられる。また、フローグラフ解析では、食材の集合 `F-set` や同一の食材 `F-eq` のラベルの認識精度が大きく向上した。そこで、提案手法は食材に関する実世界知識を多く獲得していると考え、料理オントロジーツリー [12] を用いて分析を行った。料理レシピドメインの 2 つの表現を入力し、親 (上位語) や子 (下位語) などの関係ラベルを推定するタスクを設計した。学習データとして各関係ラベルごとに 1,000 個のペアを、開発データ・テストデータとして各関係ラベルごとに 250 個のペアをランダムに選んだ。事前学習されたテキストエンコーダにそれぞれの表現を入力し、得られた 2 つの分散表現を連結した。それをフィードフォワードニューラルネットワークに入力し、関係ラベルを予測するようにファインチューニングを行った。結果を表 4 に示す。結果より、提案手法は従来手法よりも、料理に関する実世界知識を多く獲得できたと考えられる。

5 関連研究

テキストと視覚的な情報を用いて事前学習する手法が同時期にいくつか提案されており [6]、これらのモデルは画像からのキャプション生成や画像も含めた質問応答などのタスクで有用性を示している。それらと比較して、提案手法はテキスト解析タスクへの適応を目的とした。提案手法ではテキストと非テキストデータの関係性の情報が、テキスト側のエンコーダに $\boldsymbol{h}_{[\mathrm{CLS}]}$ を通して伝播する。そのため、ファインチューニング時やテスト時には、非テキストデータの入力を必要としない自然なモデルが構築でき、多くのテキスト解析タスクで有用である。

また、文献 [6] などの手法では 1 つの画像をエンコードする際に、複数の注目領域に分割する必要がある。画像をエンコードする際には、そのようなエンコーダを入手することは比較的容易だが、他の種類の非テキストデータの場合には難しいことがある、もしくは注目領域を定義することが困難な種類の非テキストデータもある。提案手法は非テキストデータエンコーダとして、任意のニューラルネットワークモデルを扱うことができる。

6 おわりに

本論文では、テキストと非テキストデータを用いて事前学習する手法を提案した。実験結果より、提案手法はテキストのみから事前学習した手法よりもテキスト解析タスクで高精度を達成することができた。また、提案手法は従来手法よりも実世界知識を多く獲得できたことを知識ベースを用いて示した。今後の方向として、文と非テキストデータの対応関係のみではなく、細かい粒度の対応関係をモデリング可能な手法の提案などが挙げられる。

参考文献

[1] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

[3] J. Harashima, Y. Someya, and Y. Kikuta. Cookpad image dataset: An image collection as infrastructure for food research. In *ACM SIGIR*, pp. 1229–1232, 2017.

[4] G. Jawahar, B. Sagot, and D. Seddah. What does BERT learn about the structure of language? In *ACL*, pp. 3651–3657, 2019.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.

[6] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[7] Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. A framework for procedural text understanding. In *IWPT*, pp. 50–60, 2015.

[8] S. Matsuyoshi, H. Kameko, Y. Murawaki, and S. Mori. Annotating modality expressions and event factuality for a Japanese chess commentary corpus. In *LREC*, 2018.

[9] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. Non-projective dependency parsing using spanning tree algorithms. In *EMNLP*, pp. 523–530, 2005.

[10] S. Mori, H. Maeta, Y. Yamakata, and T. Sasada. Flow graph corpus from recipe texts. In *LREC*, pp. 2370–2377, 2014.

[11] S. Mori, J. Richardson, A. Ushiku, T. Sasada, H. Kameko, and Y. Tsuruoka. A Japanese chess commentary corpus. In *LREC*, 2016.

[12] H. Nanba, Y. Doi, M. Tsujita, T. Takezawa, and K. Sumiya. Construction of a cooking ontology from cooking recipes and patents. In *ACM UBICOMP*, pp. 507–516, 2014.

[13] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *EMNLP*, pp. 2463–2473, 2019.

[14] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.