BERTを用いたバイナリパタン間の含意関係認識

二宮 大空[†‡] 門脇 一真[‡§] 飯田 龍[†‡] 鳥澤 健太郎[†‡] Julien Kloetzer[‡]
[†]奈良先端科学技術大学院大学 先端科学技術研究科
[‡]国立研究開発法人 情報通信研究機構
[§]株式会社日本総合研究所
ninomiya.hirotaka.ng8@is.naist.jp,
{kadowaki, ryu.iida, torisawa, julien}@nict.go.jp

1 はじめに

本研究では「XはYの首都である」と「XがYにある」のようなパタンの間に含意関係が成り立つか否かを認識する含意関係認識の問題を扱う．一般に，含意関係認識の問題は「夏目漱石は夢十夜の著者である」と「夏目漱石が夢十夜を執筆した」のような2つの文の間の含意関係認識を対象としており，盛んに研究が進められている [1][2][3]．一方，本研究で扱うパタン間の含意関係認識では,「XがYを執筆した」のようにKloetzerら [4][5] が対象とした2つの変数を持つバイナリパタンを対象に扱う．このバイナリパタン間の含意関係認識処理は，大規模なWeb文書を対象としたファクトイド型質問応答で重要であり，実際に大規模情報分析システムWISDOM X[1][6] 等で既に活用されている．例えば，質問文「地球温暖化で何が起きる？」に対して，(1) 質問文中の名詞「地球温暖化」と (2) 質問文から得られるバイナリパタン「XでYが起きる」，(3)(2) と含意関係が成り立つバイナリパタン「XでYが発生する」を用いて検索することで,「地球温暖化で異常気象が発生する」といった文から答え「異常気象」を抽出できる．

最近ではBERT（Bidirectional Encoder Representations from Transformers）[7] を利用した含意関係認識も盛んに研究が進められている [8][9]．しかし，一般にBERTへの入力は自然な文が想定されており，バイナリパタンのペアやそれに付随する情報をどのように入力すれば高精度に含意関係が認識できるかは自明ではない．そこで，本研究ではバイナリパタン間の含意関係認識において4種類の入力形式を提案し，それらの有効性の調査を行った．実験の結果，提案する入力形式を組み合わせることで，Kloetzerら [5] の性能を

表 1: バイナリパタン間の含意関係認識の具体例

含意関係が成り立つバイナリパタンペア		
•	X は Y の 首都 で ある	→ X が Y に ある
•	X の 新しい Y	→ X の Y
含意関係が成り立たないバイナリパタンペア		
•	X を 保証 する Y	↛ X を 低下 さ せる Y
•	X まで Y を 読む	↛ X から Y を 読む

2つのバイナリパタン P, Q について P が Q を含意している場合を $P \rightarrow Q$ と表し，含意しない場合を $P \nrightarrow Q$ と表す.

F値，平均精度でそれぞれ約24ポイント，21ポイント上回った．Kloetzerら [5] がWeb6億ページから獲得した含意関係の成り立つバイナリパタンのペア（以下，含意パタンペア）は精度80%で約2.2億件であるのに対して，本研究で構築した分類器を用いた場合はその精度を保ったままWeb6億ページから約5.6億件を獲得できる見込みである．

2 関連研究

「XはYの首都である」,「Xの新しいY」,「XをYで食べる」のように係り受けで繋がれた2つの名詞を変数に置換したパタンをバイナリパタンという．ここで,「XがYの首都である」ことからは「XがYにある」ことがわかるため，前者は後者を含意している．本研究ではこのようなバイナリパタン間の含意関係認識を行う．その具体例を表1に示す．

バイナリパタン間の含意関係認識に関する研究として，Linら [10] の教師なしのスコアリング手法が挙げられる．一方，Kloetzerら [4][5] は，教師あり学習データを用いて，バイナリパタン間に含意関係が成立するか否かを分類するSVMベースの分類器を学習させた．さらに，この分類器をWeb6億ページから抽出したバイナリパタンペアに適用し，約2.2億件の含意パタンペアを獲得した．Kloetzerらが利用した含意関係認識

[1]https://www.wisdom-nict.jp

表 2: 提案する入力形式のフォーマットと具体例

Pattern	フォーマット	P [SEP] Q
	具体例	X は Y の 首都 で ある [SEP] X が Y に ある
PatClassWord (PCW)	フォーマット	P [SEP] Q [SEP] C [D] C' [SEP] N_1 [D] N_1' [D'] N_2 [D] N_2' [D'] $\cdots$ [D'] N_T [D] N_T'
	具体例	X は Y の 首都 で ある [SEP] X が Y にある [SEP] C100 [D] C200 [SEP] 東京 [D] 日本 [D'] 北京 [D] 中国
FillClass (FC)	フォーマット	P^C [SEP] Q^C
	具体例	C100 は C200 の 首都 で ある [SEP] C100 が C200 に ある
FillWord (FW)	フォーマット	P_1^W [D] Q_1^W [D'] P_2^W [D] Q_2^W [D'] $\cdots$ [D'] P_T^W [D] Q_T^W
	具体例	東京 は 日本 の 首都 で ある [D] 東京 が 日本 に ある [D'] 北京 は 中国 の 首都 で ある [D] 北京 が 中国 に ある

バイナリパタンペアを P，Q，二変数に対する単語意味クラスを C，C'，$T(1 \leq T \leq 9)$ 個ある名詞ペアの内 t 番目の名詞ペアを N_t，$N_t'(1 \leq t \leq T)$，シーケンスを区切るために導入した事前学習のデータには出現しないトークンを [D]，[D'] とする．二変数を単語意味クラスで置換したバイナリパタンペアを P^C，Q^C とし，二変数を $t(1 \leq t \leq T)$ 番目の名詞ペアに置換したバイナリパタンペアを P_t^W，Q_t^W とする．未知語はトークン [UNK] とし，トークン [SEP] で segment embedding を変更した．具体例は東京と北京の単語意味クラスが 100，日本と中国の単語意味クラスが 200 の場合である．

データは (1) バイナリパタンペア，(2) バイナリパタン中の変数 X，Y に対する単語の意味を表すクラス（以下，単語意味クラス），(3) バイナリパタン中の変数に置換前の最大 9 組の名詞のペア，(4)ALAGIN 言語資源[2]（動詞含意関係データベース，日本語異表記対データベース，基本的意味関係の事例ベース）に含まれているかどうかを表す素性で構成されている．(2) の単語意味クラスは，Kazama ら [11] が行った，単語間の係り受け関係のクラスタリングの結果を利用して 100 万の名詞に 500 種の単語意味クラスを割り当てたものである．

本研究では Kloetzer ら [5] の手法をベースラインとし，同じデータを用いて実験を行う．ただし，Kloetzer らが用いた ALAGIN 言語資源は単語や動詞句に関するデータであり，提案する BERT ベースのモデルでの利用方法が自明でないため，本研究の実験では利用しない．

3 提案する BERT への入力形式

BERT は入力として自然な文を想定しており，バイナリパタンといった通常の入力文中には存在しないシーケンスを想定していない．そこで，バイナリパタンペア，二変数，単語意味クラス，バイナリパタン中の二変数に対応する名詞ペア集合を組み合わせた 4 種類の BERT への入力形式を提案する．データの作成には Kloetzer ら [5] のデータを用いており，500 種類存在する単語意味クラスは C1 から C500 の単語で表す．提案する 4 種類の入力形式は以下の通りである．

Pattern

バイナリパタンペアのみの形式

PatClassWord(PCW)

バイナリパタンペアの後に，単語意味クラスと名詞ペア集合を列挙した形式

FillClass(FC)

バイナリパタンペアの二変数を対応する単語意味クラスに置換した形式

FillWord(FW)

バイナリパタンペアの二変数を対応する名詞ペアに置換したものを列挙した形式

表 2 にこれらのフォーマットと具体例を示す．

さらに，各形式に変換された入力データごとに，BERT が異なる特徴を学習すると考えたため，Pattern 以外の各形式のシーケンスをトークン [SEP] を用いて組み合わせることで，4 種類の入力形式 PCW+FC，PCW+FW，FC+FW，PCW+FC+FW のデータを作成した．これらはそれぞれ PatClassWord と FillClass，PatClassWord と FillWord，FillClass と FillWord，PatClassWord と FillClass と FillWord を組み合わせたデータの形式を表している．

Train データ，Dev データ，Test データにおけるデータセットの件数，正例の件数と全体に対する正例の割合を表 3 に示す．これらは入力形式に依らず，全て一定である．

表 3: データセットの統計

	件数	正例	正例割合
Train	80,107	23,722	29.6%
Dev	5,000	430	8.6%
Test	15,000	1,326	8.8%

[2]https://alaginrc.nict.go.jp

表 4: 実験結果

入力形式	モデル	再現率	精度	F 値	平均精度
	Kloetzer et al.[5]	29.79	**65.94**	41.04*†	50.64
Pattern	$\text{BERT}_{\text{BASE}}$	53.17	56.54	54.80*†	58.34
PatClassWord	$\text{BERT}_{\text{BASE}}$	58.60	60.94	59.75*†	64.50
FillClass	$\text{BERT}_{\text{BASE}}$	54.30	57.37	55.79*†	60.68
FillWord	$\text{BERT}_{\text{BASE}}$	59.43	63.91	61.59†	66.77
PCW+FC	$\text{BERT}_{\text{BASE}}$	56.18	60.67	58.34*†	63.26
PCW+FW	$\text{BERT}_{\text{BASE}}$	61.46	59.45	60.44*†	67.07
FC+FW	$\text{BERT}_{\text{BASE}}$	58.30	65.73	61.79	67.44
PCW+FC+FW	$\text{BERT}_{\text{BASE}}$	58.22	62.71	60.38*†	66.12
FC+FW	$\text{BERT}_{\text{LARGE}}$	**63.65**	65.89	**64.75**	**71.87**

入力形式 FC+FW の $\text{BERT}_{\text{BASE}}$ モデルと比べて FC+FW よりも F 値が低く，マクネマー検定（有意水準 5%）によって有意差が確認された場合は F 値に * で示す．入力形式 FC+FW の $\text{BERT}_{\text{LARGE}}$ モデルと比べて FC+FW よりも F 値が低く，マクネマー検定（有意水準 5%）によって有意差が確認された場合は F 値に † で示す．

4 実験

本節では 3 節で述べた含意関係認識データを用いて，提案する BERT の入力形式の有効性を調査する．

4.1 実験設定

実験で利用したモデル設定は Devlin ら [7] の $\text{BERT}_{\text{BASE}}$，$\text{BERT}_{\text{LARGE}}$ に従う．$\text{BERT}_{\text{BASE}}$ の事前学習には，Kadowaki ら [12] の実験と同様に因果関係を含むテキスト約 2,000 万文のコーパスを用いる．このコーパスは，7 文からなるパッセージの集合で構成されており，それらは Oh ら [13] の因果関係抽出器により検出された文とその前後の文で構成された 7 文である．一方，$\text{BERT}_{\text{LARGE}}$ は同様の手法で獲得した約 12 億文のコーパスを用いる，バッチサイズはいずれのモデルも Kadowaki らと異なり 4,096 である．fine-tuning 時のバッチサイズは 32 とし，Dev データを用いたパラメータ探索によって F 値，平均精度が最大となる学習率とエポックをそれぞれ {1e-5，2e-5，3e-5，4e-5，5e-5} と {1，2，3} から選択した．

4.2 実験結果と分析

BERT の二値分類における閾値を 0.5 として算出した，Test データにおける再現率，精度，F 値と平均精度を表 4 に示す．モデル $\text{BERT}_{\text{BASE}}$ において F 値と平均精度は FC+FW が最高であるため，$\text{BERT}_{\text{LARGE}}$ は FC+FW に対してのみ実行した．表 5 に Dev データ（正例）の具体例を示す．

表 4 において Pattern，PatClassWord，FillClass，FillWord を比較すると，Pattern よりも他 3 つの入力形式を用いた場合の F 値，平均精度が高いことがわかる．そのため，単語意味クラスや名詞ペア集合といったバイナリパタンペアに付随する情報は BERT でトークン列として入力した場合でも有効であることがわかる．さらに，FillWord は単語意味クラスの情報を含まないにも関わらず，単語意味クラスの情報を含む PatClassWord よりも F 値，平均精度が高い．2 つの形式を比較すると，PatClassWord はバイナリパタンペア，単語意味クラス，名詞ペア集合を断片的に入力しているのに対し，FillWord はバイナリパタン中の二変数を名詞ペアに置換することで自然な文に近い形式で入力している．つまり，FillWord の方が BERT の事前学習時の入力に近い形式で入力しているために，PatClassWord よりも分類性能が高いと考えられる．

また，FillWord は PatClassWord，FillClass どちらに組み合わせてもスコアが上昇した．特に，FC+FW は F 値，平均精度において，Kloetzer et al. を約 21 ポイント，17 ポイント，PatClassWord を約 2 ポイント，3 ポイントそれぞれ上回った．さらに，入力形式 FC+FW でモデルを $\text{BERT}_{\text{BASE}}$ から $\text{BERT}_{\text{LARGE}}$ に変更することで，F 値，平均精度はそれぞれ約 3 ポイント，4 ポイント上昇した．

最後に，Test データの分類結果を用いて図 1 に再現率-精度曲線を図示する．この結果から，精度 80%の場合，入力形式 FC+FW でモデル $\text{BERT}_{\text{LARGE}}$ を用いた場合は Kloetzer et al. の 2 倍以上の含意パタンペアを獲得できることがわかる．

表 5: Dev データ（正例）の具体例

入力データ			分類スコア	
バイナリパタンペア	単語意味クラス	名詞ペア	Pattern	FC+FW
X は Y を 求めて いる → X は Y に ついて 考えて いる	X=C274 Y=C269	X_1 = オレ /Y_1 = 成功，X_2 = 私 / Y_2 = 非 日常 生活，X_3 = 僕 / Y_3 = 人生…	0.302	0.974
Y に は X が 植えて ある → Y に は X が 生えて おる	X=C227 Y=C494	X_1 = 木 / Y_1 = 北側，X_2 = 桜 / Y_2 = 中腹，X_3 = 木 / Y_3 = 境内…	0.166	0.903
Y で X まで 送って くれた のだ → X は Y で 移動 だ	X=C358 Y=C132	X_1 = 自宅 / Y_1 = 車，X_2 = ホテル / Y_2 = マイカー，X_3 = 駅 / X_3 = バス…	0.074	0.537

分類スコアは入力形式 Pattern，FC+FW を用いた場合の $\text{BERT}_{\text{BASE}}$ の出力を表す．

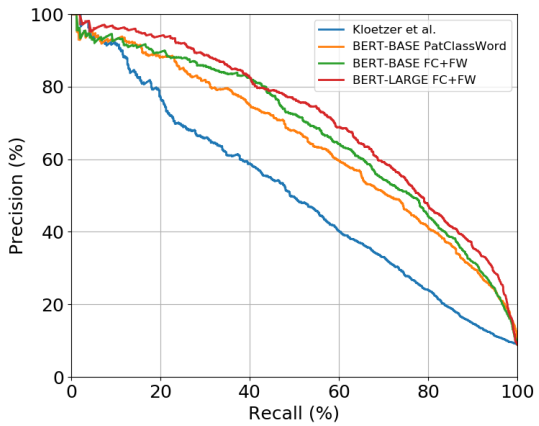

図 1: 再現率-精度曲線

5 おわりに

本研究ではバイナリパタン間の含意関係認識において，BERT の入力とするデータの形式を 4 種類提案した．実験の結果，単純に付随する情報をつなげた形式と比べて，バイナリパタン中の二変数を対応する名詞に変換することでより自然な文に近づけた形式は F 値，平均精度において，それぞれ約 2 ポイント，3 ポイント高い分類性能となった．さらに，単語意味クラスに変換した形式と組み合わせた形式 FC+FW を用いた $\text{BERT}_{\text{LARGE}}$ の分類性能が最高となり，Kloetzer ら [5] の性能を F 値，平均精度において，それぞれ約 24 ポイント，21 ポイント上回った．Kloetzer ら [5] が Web6 億ページから獲得した含意関係の成り立つバイナリパタンのペアは精度 80%で約 2.2 億件であるのに対して，本研究で構築した分類器を用いた場合はその精度を保ったまま Web6 億ページから約 5.6 億件を獲得できる見込みである．

今後は，本研究のモデルを用いて含意関係の成り立つバイナリパタンのペアの抽出を大規模に行うことを考えている．

参考文献

[1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pp. 177–190, 2005.

[2] Jonathan Berant, Ido Dagan, and Jacob Goldberger. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 610–619, 2011.

[3] Daniel Bikel and Imed Zitouni. *Multilingual natural language processing applications: from theory to practice*. IBM Press, 2012.

[4] Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Motoki Sano, Jun Goto, Chikara Hashimoto, and Jong-Hoon Oh. Supervised recognition of entailment between patterns. 言語処理学会第 18 回年次大会講演論文集, pp. 431–434, 2012.

[5] Julien Kloetzer, Kentaro Torisawa, Chikara Hashimoto, and Jong-Hoon Oh. Large-scale acquisition of entailment pattern pairs by exploiting transitivity. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1649–1655, 2015.

[6] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 263–267, 2016.

[7] 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*, pp. 4171–4186, 2019.

[8] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, 2019.

[9] Noha Tawfik and Marco Spruit. UU_TAILS at MEDIQA 2019: Learning textual entailment in the medical domain. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 493–499. Association for Computational Linguistics, 2019.

[10] Dekang Lin and Patrick Pantel. DIRT – discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 323–328, 2001.

[11] Jun'ichi Kazama and Kentaro Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, pp. 407–415, 2008.

[12] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5815–5821, 2019.

[13] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1733–1743, 2013.