BERTによる英語テキストからの因果関係抽出

矢野 憲　　奥村 学
東京工業大学 科学技術創成研究院
{yano, oku}@lr.pi.titech.ac.jp

1 はじめに

本研究では，ニュース記事などの英語テキストから因果関係を自動で抽出する手法を提案する．因果関係の認識は，質問応答，情報検索，知識学習などのアプリケーションで特に注目されている．例えば，WWWなどのオンラインコンテンツや，オンラインジャーナル等から取得した因果関係を用いて因果連鎖を構成することができれば，以前に知られていなかったエンティティ間の因果を発見することも可能となる．因果関係は，医学や生物学で特に重要であるが，金融や経済分野においても企業や投資家の意思決定・政策決定において重要な判断材料となりうる．

英語テキストの因果関係は手がかり表現を伴い明示的に表現される場合もあるが，曖昧で非明示的に表現されることも非常に多い．また因果関係は文内，文間のいずれの場合も存在する．このため，テキスト中から高い精度で因果関係を抽出することは一般的には非常に困難なタスクであると認識されている．

一方で，最近の自然言語処理では，大量のテキストデータを用いて学習した大規模言語処理モデルを転移学習により比較的少量の教師データが利用できる他の言語処理タスクへ適用し，これまでよりも性能が飛躍的に向上することが数々の研究で示されている．本研究でもこの手法を用いて，BERT[2] による事前学習モデルを利用してテキストに現れる因果関係を自動抽出する手法の提案と評価を行なった．本研究では，因果関係抽出のコーパスとして，Penn Discourse Treebank(PDTB) 3.0 [6] を用いた．

2 関連研究

英語テキストにおける因果関係は様々な形式で表現・定義されている．例えば，因果を関係づける接続詞または動詞により表現される場合や複合名詞句の内部構造で表現される場合がある [3]．例えば，SemEval-2010 Task8[4] では，文中の名詞語句間の因果関係認識のタスクを定義している．

一方．PDTB のコーパスでは，明示的または非明示的に隣り合う文，節またはフレーズ間で定義される因果関係をアノテーションしているため，本研究では，PDTB で定義される低レベルの談話構造による因果関係の抽出を目的とする．

本研究は，一般的な PDTB パーサーや PDTB を用いた談話関係クラス認識にも関連している [7]．つまり因果だけでなくその他の談話関係ラベルも分類できるように拡張すれば，提案した手法を一般的な PDTB パーサーや談話関係認識タスクへの応用も可能である．PDTB パーサーの研究はこれまでにも，構文木を利用した方法 [5] や，タグ付けの手法を用いた研究例 [1] がある．

3 学習データセットの構築

因果関係の教師データとして，本研究では Penn Discourse TreeBank(PDTB) 3.0 を使用した．PDTB は WSJ(Wall Stree Journal) コーパスに対して，文内や文間における浅いレベルでの談話構造のラベル付けを行なったコーパスである．PDTB 3.0 は旧来の PDTB 2.0 に比べて量，質ともに向上していると述べられている [6]．PDTB のタグ付けは，基本的には手がかりとなる単語やフレーズを用いて関係がタグ付けされる場合と，そのような手がかり表現が無く，暗黙的に隣り合う文またはフレーズ間でラベル付けされる 2 つのパターンが存在する．談話関係のラベル（PDTB では relation sense と表現される）は 3 段階の階層構造で定義されており，一番上位レベルの Level-1 では Temporal, Contingency, Comparison, Expansion が定義され，それぞれの下位レベル Level-2,Level-3 でより詳細な談話関係ラベルが定義されている．

本研究では，この中で特に因果関係を定義している Level-2 の Contingency.Cause のラベルでタグ付けさ

れているものを陽性 (Pos) のサンプル，それ以外を陰性 (Neg) のサンプルとして，学習データを構築した．

コーパス中に存在する Contingecy.Cause の例を以下に示す．

- Runways at San Francisco weren ' t damaged, *but traffic was being limited yesterday to 27 arrivals and 27 departures an hour – down from 33 to 45 an hour normally* – mainly because **the noise level in the control tower was overwhelming without the windows**, an FAA spokeswoman said. [wsj 1803]
- *But service on the line is expected to resume by noon today.* (Implicit=since) **"We had no serious damage on the railroad**," said a Southern Pacific spokesman. [wsj 1803]
- *Now, though, enormous costs for earthquake relief will pile on top of outstanding costs for hurricane relief.* "That obviously means that **we won ' t have enough for all of the emergencies that are now facing us**, and we will have to consider appropriate requests for follow-on funding," Mr. Fitzwater said. [wsj 1824]

例でイタリック，太字，アンダーラインで表記されているものがそれぞれ ARG1,ARG2 そして手がかり表現となる CONN(discourse connective) を示している．最初の 2 つの例は Contingency.Cause.Reason としてラベル付けされており，ARG2 が理由，説明，根拠となり ARG1 が結果となる場合で，最初が CONN が存在する Explicit の場合で，2 番目は CONN が存在しない Implicit の場合である．最後の例が Contingency.Cause.Result の場合で，ARG1 が理由，説明，根拠となり ARG2 が結果の場合の例である．なお，PDTB では文の構文上 CONN と直接関係するフレーズを ARG2，もう片方を ARG1 と決めている．

本研究では，ARG1,ARG2,CONN で定義されるテキストスパン（区間）を NER（固有表現認識）で使用される IOB2 フォーマットに置き換えて学習データセットを構築した．

4 提案する手法

本研究では，因果関係の抽出を，ARG1,ARG2,CONN ラベルのスパン認識のタスク及び，ARG1 と ARG2 で示されるスパン間の関係が因果関係であるか否かの二値分類タスクとに区別する．提案手法は，BERT による事前学習モデルを用いてラベルのスパン認識を行なった後，入力テキストと推測されたタグのスパン情報を補助的な入力として双方向の GRU を用いて因果関係ラベルの分類を行なった．

提案手法はこのように因果関係抽出のタスクを 2 段階で行うが．ラベルのスパン認識と同時に関係ラベルの分類も BERT モデルの出力を利用して学習する手法も考えらえる．しかし実験の結果，提案する 2 段階での推論モデルの方が，BERT を利用したマルチタスクモデルよりも性能が良いことが示された．詳細については結果で述べる．

図 1 に提案モデルの概要を示した．入力 (input) は，1 文または連続する 2 文であり，入力の先頭に "[CLS]"，文の区切りに "[SEP]"のタグを埋め込んで BERT が求める入力フォーマットに変換を行う．次に，個々の入力トークンを BERT が持っている語彙変換辞書により ID 変換して BERT モデルへ入力する．BERT の出力は全結合層を介して CRF への入力となり，教師データであるタグ情報を用いて系列学習を行う．学習時の Loss は CRF 層の Loss を用いた．タグは ARG1, ARG2, CONN の 3 種類存在するため，用いた系列ラベルは B-ARG1, I-ARG1, B-ARG2, I-ARG2, B-CONN, I-CONN, O の 7 種類である．

図 1 の右側は BERT+CRF により推測されたタグ情報と元の入力を用いて双方向の GRU により因果関係クラスの分類を行う．GRU への入力は，入力トークンの Embedding と該当するタグの Embedding を連結したベクトルとなる．GRU は双方向の一層で構成され，各方向の最終の隠れユニットを連結したベクトルを MLP を等して，シグモイド関数による 2 値分類を行った．学習時の Loss はクロスエントロピーを用いた．なお，GRU の学習時のタグ情報は教師タグを用いて学習を行なった．

表 1 に BERT モデル，GRU への入力の語彙数とその分散表現のサイズを示した．GRU の隠れユニットのサイズは 100 とした．使用した学習済みの BERT モデルは 12 層で構成され，Bert の語彙変換辞書は大文字・小文字の区別があるものを使用した[1]．

各々のモデルは別々に学習した．いずれのモデルも学習時のバッチサイズは 16，エポック数は 50 で行った．最適化には Adam を用いて，学習率は 0.005 から

[1] https://huggingface.co/transformers/

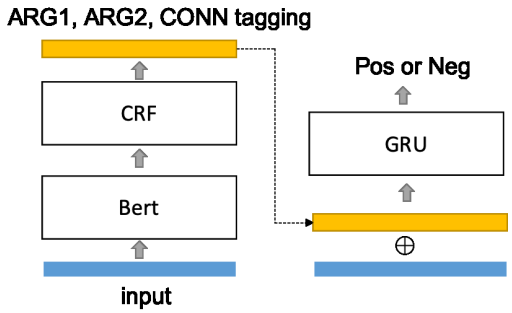

図 1: 提案モデルの概要. 左:入力文から ARG1, ARG2, CONN 系列タグ認識を行うモデル. 右：推測されたタグ系列と入力文を用いて因果関係クラスの認識を行うモデル

	語彙サイズ	分散表現サイズ
BERT 入力トークン Embed	28996	768
GRU 入力トークン Embed	28996	200
GRU 入力タグ Embed	7	100

表 1: BERT, GRU の入力語彙サイズとその分散表現サイズ

減少率 0.5 で徐々に学習率を小さくして行った．評価には，Validation データによる性能が最も高かったパラメータを用いて行った．

5 結果

本研究では，PDTB 3.0 に含まれる 0 から 24 までのデータセットで，最初のセットを Validation，最後のセットを Test，それ以外を Train として学習データを構築した．表 2 に Train, Val, Test データセットの Pos,Neg ラベル毎のサンプル数と，その PDTB 内での関係タイプの内訳を示した．因果関係の有り無しを区別する Pos,Neg ラベルのサンプル数は Train データセットでは Neg の方が Pos よりも約 5 倍多いため，2 値ラベルのクロスエントロピーの計算では Pos ラベルのサンプルに 5.0 の重み付けを行なって学習し評価を行なった．

まず，BERT+CRF によるタグの識別結果を表 3 に示す．タグ識別の性能は NER の性能評価と同様の手法で計算を行なった．次に，元の入力テキスト及び BERT+CRF で推測されたタグ情報を GRU への入力として，因果関係の 2 値分類を行った結果を表 4 の上段に示した．結果からわかるように Neg ラベルの性能は F1 が 85.85 と高いものの，Pos ラベルの F1 は 32.97 とかなり低い結果となった．Pos と Neg の結果の F1 のマイクロ平均は 76.63, マクロ平均は 59.40 であった．参考までに，GRU へ入力するタグ情報を BERT+CRF で推測されたものでなく，教師で与えられたタグ情報に置き換えた場合の結果を表 4 の下段に示した．提案手法の結果と比べて Pos の全てのスコアが約 10 のオーダーで上昇している，これは正確なタグのスパン認識が因果関係の認識の性能に非常に影響していることを意味している．

比較のために，BERT モデルを利用して ARG1,ARG2,CONN の系列タグ認識と因果関係ラベルの分類を同時に行った場合の結果を表 5 に示す．なお，ラベル分類については，BERT の出力を入力テキストのシーケンス方向に平均したベクトルを 3 層の MLP(隠れユニット数が 256,64,32)，第 1,2 層の activation を ReLU，最後にシグモイド関数による 2 値分類とした．この場合の F1 のマイクロ平均は 64.30 (-12.23)，マクロ平均は 58.56 (-0.84) であった．カッコ内の数字は，GRU によりラベル分類を行った結果との比較である．平均としては提案手法の性能の方が高いが，BERT で同時学習した場合の Pos ラベルの Recall が 76.81 と提案手法と比べて非常に大きな値を示した．これは BERT によるラベル認識では，より Pos ラベルを付与しやすくなった結果だと言える．

6 考察

表 6 にテストデータにおける ARG1，ARG2 の関係種別毎の提案手法の Pos，Neg ラベルの Precision,Recall,F1 平均を示した．CONN が存在する Explicit では，F1 のマイクロ平均で 85.48 とかなり高い性能であることが確認できる．一方で Implicit や AltLex,AltLexC の場合の性能は，Explicit の場合と比べてかなり低いと言える．ただし，Pos ラベルに限れば，Explicit の場合でも性能はそれほど高くない．この原因としては，Train データセット中の Pos サンプルが十分でないことも一因であると考えられる．

7 おわりに

本稿では，BERT の事前学習モデルを用いてニュース記事などの英文テキストからの因果関係を抽出する手法を提案した．学習データには PDTB 3.0 を用い，

データセット	ラベル	サンプル数	ARG1 と ARG2 の関係タイプの内訳					
			Explicit	Implicit	AtlLex	AltLexC	EntRel	Hypophora
Train	Pos	8368	2127	5360	854	27	-	-
	Neg	41011	19907	14917	575	96	5141	127
Val	Pos	327	327	97	199	31	-	-
	Neg	1759	834	634	19	6	230	4
Test	Pos	276	73	182	20	1	-	-
	Neg	1290	163	419	25	10	163	9

表 2: Train, Val, Test データセットにおけるラベル毎のサンプル数と関係タイプの PDTB での内訳

	Precision	Recall	F1
ARG1	47.72	40.97	44.09
ARG2	53.60	48.63	50.99
CONN	58.22	58.78	58.50

表 3: BERT+CRF モデルによる ARG1, ARG2, CONN のタグの識別性能の結果

	ラベル	Precision	Recall	F1
提案手法	Pos	33.33	32.61	32.97
	Neg	85.65	86.05	85.85
教師タグ	Pos	42.20	43.12	42.65
を用いた場合	Neg	87.77	87.36	87.57

表 4: 提案手法による Pos, Neg 因果関係ラベル識別の性能結果と教師タグ情報を用いた場合の性能との比較

この中で因果に関連するサンプルを Pos，それ以外を Neg とみなして学習データを構築した．提案モデルは，1 文または連続する 2 文で，最初にペアとなるフレーズかつ手がかりとなる単語もしくはフレーズを BERT モデルを利用したタグ付けを行い，その結果を利用して GRU により因果関係の識別を行う 2 段階の推論方式を提案した．

今後の課題としては，さらなる性能の向上と実データでの性能評価を行う予定である．

8 謝辞

本研究は，三菱 UFJ 銀行と東京工業大学との共同研究の補助によって行われた．

参考文献

[1] Or Biran and Kathleen McKeown. PDTB discourse parsing as a tagging task: The two taggers approach. In *Proc. of SIGDIAL*, pp. 96–104, September 2015.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.

[3] Roxana Girju and Dan Moldovan. Mining answers for causation questions. In *In AAAI symposium on*, 2002.

[4] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, July 2010.

[5] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, Vol. 20, pp. 151–184, 2010.

[6] Prasad Rashmi, Webber Bonnie, Lee Alan, and Joshi Aravind. *The Penn Discourse Treebank 3.0*. Linguistic Data Consortium, University of Pennsylvania, 3 2019.

[7] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol T Rutherford. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. Technical report.

	Precision	Recall	F1
Pos	29.99	76.81	43.13
Neg	92.55	61.63	73.99

表 5: BERT モデルで ARG1,ARG2,CONN の系列認識及び，Pos, Neg 因果関係ラベル識別を同時に行った場合の因果関係認識の結果

		Explicit	Implicit	AltLex AltLexC
P	Pos	34.29	37.40	35.71
	Neg	94.08	71.70	61.90
R	Pos	49.31	26.92	23.81
	Neg	89.50	80.43	74.29
F1	マイクロ平均	85.48	64.22	55.36
	マクロ平均	66.09	53.56	48.05

表 6: ARG1,ARG2 の関係種別毎の提案手法の Pos,Neg ラベルの Precision(P), Recall(R), F1 平均