The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

日本語BERTモデルを用いた 経済テキストデータのセンチメント分析

Sentiment Analysis of Japan Economic Watcher Survey Data with Japanese BERT model

青嶋 智久 *1*3　中川 慧 *2
Tomohisa Aoshima　Kei Nakagawa

*1富士通クラウドテクノロジーズ株式会社　*2野村アセットマネジメント株式会社
Fujitsu Cloud Technologies Limited.　Nomura Asset Management Co,ltd

*3筑波大学大学院 リスク工学専攻
Department of Risk Engineering, University of Tsukuba

A new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers obtains new state-of-the-art results on eleven natural language processing tasks in English. We build a Japanese version of BERT model with Japanese Wikipedia data and perform sentiment analysis of Japan Economic Watcher Survey Data. We confirmed that the result of sentiment analysis using the Japanese version of BERT model is better than the result without the model.

1. はじめに

政府や日銀、あるいは様々な金融機関が、膨大な数の経済レポートを日々発行している。投資家はこれらの膨大なレポートから投資行動に有用な情報を得るために、非常に多くの時間を費やす必要がある。しかしながら投資家は時間的な制約から、膨大なレポートの一部しか読むことが出来ず、大量のテキストデータを十分に活用出来てない。そういった問題意識から、経済テキストデータをコンピューターが自動処理し、投資判断に活用する金融テキストマイニング技術が金融市場で普及してきた。金融テキストマイニングでは、あるテキストが表す状況が市場や景気に対してポジティブもしくはネガティブかを評価するセンチメント分析の事例が多くある [和泉 17]。

センチメント分析には様々な手法が存在し、[Turney 02] などの極性辞書ベース・アプローチと [Pang 02] などの機械学習アプローチとに大別される。極性辞書ベース・アプローチでは、テキスト中に出現するポジティブな単語 (市場や景気にとって良い意味の単語) の出現比率とネガティブな単語の出現比率の差で対象となるテキスト全体の極性を算出する方法が用いられる。一方、機械学習アプローチでは、テキストの特徴量とポジティブやネガティブなどの極性ラベルとの関係性を、機械学習モデルによって学習し、極性付与されていないテキストに対して学習済みモデルを適用することで極性を付与する。このアプローチで特に、深層学習を用いることで高い精度が得られることが報告されている。例えば、RNN を用いた金融関連テキストの分析としては、ニュース記事の要約を試みた [Chen 15] がある。また LSTM を用いて景気ウォッチャー調査 [内閣] を学習データに用いることで、経済テキストのセンチメント分析を行った [山本 16] がある。

近年、[Devlin 18] は大規模なコーパスによる双方向の Transformer [Vaswani 17] を用いた教師なし学習による事前学習を行い、その事前モデル (BERT モデル) を個別のタスクごとにファインチューニングすることで自然言語処理の様々なタスクにおいて最高のスコアを更新した。

本研究では、[Devlin 18] の BERT モデルを Wikipedia 日本語版データ [Wikimedia Foundation] を用いて学習し、その事前モデルを用いて内閣府の景気ウォッチャー指標を学習したセンチメント分析を行う。そして、[山本 16] よりも高精度のセンチメント分析が可能であることを実証する。

連絡先: 青嶋 智久,
富士通クラウドテクノロジーズ株式会社
筑波大学大学院 リスク工学専攻
tomohisa.aoshima@blue271828.com

2. BERT モデル

様々な自然言語処理タスクは、トークンレベルのタスク、文レベルのタスクに大別できる。トークンレベルのタスクはトークン同士の関係性を学ぶのに対して、文レベルのタスクは文同士の関係性を学ぶ。いずれのタスクにおいても、事前学習は有効であることが確認されている。様々なタスクへの応用を前提とし、有効な特徴量を学習されることを事前学習という。事前学習のアプローチは特徴量ベースのアプローチとファイン・チューニングのアプローチがある。前者は自然言語処理において代表的な N-gram モデルや Word2Vec [Mikolov 13] などが該当し、事前学習により得られた特徴量をそのまま、様々な NLP タスクに利用する。後者は、次の単語を予測する言語モデルを事前学習として行い、その後にタスクに応じた教師あり学習でファイン・チューニングするアプローチである。BERT モデルは後者のファイン・チューニングアプローチによる事前学習モデルで、具体的には双方向 Transformer を用いたモデルである。

Transformer は、Attention のみを使用したニューラル翻訳モデルで、わずかな訓練で LSTM よりも良好な結果を得ている。Attention とは、系列データを扱う際に重要な箇所に注目 (Attention) するための手法で、加法注意・内積注意・自己注意等がある。中でも、自己注意 (Self-Attention) は汎用的で有効であり、言語処理のあらゆるタスクにおいて高い性能を発揮することが知られている [Vaswani 17]。BERT モデルはこの Transformer をユニットとして複数個利用している。

BERT モデルの特徴のひとつは、注目する単語の前後の文脈の両方 (双方向) について学習している点にある。従来のモデルでは単語の先読みをしてしまうという問題から前の文脈 (単方向) しか利用できなかったのに対して、BERT モデルは学習方法を事前学習のタスクを工夫することで、双方向での学

習が可能となった。BERT モデルにおける事前学習のタスクは以下の 2 つである。

1. マスクされた単語の予測

   N 個の単語を持つ文章において、$n-1$ 番目までの単語から n 番目の単語を予測するタスクにおいては、予測モデルにおいて予測するべき未来の単語情報 (n 番目以降の単語) を予測に用いないような制約が先読みを防ぐために必要となる。したがって、モデル自身は後の文脈を考慮できず、前の文脈のみで次の単語を予測する必要があった。しかし、モデル側ではなく入力データをマスク化し制約をかけることことで、前後の文脈を考慮するモデルで学習が可能となった。具体的には、 系列の 15%を [MASK] トークンに置き換えて予測を行う。[MASK] トークンのうち、80%がマスク、10%がランダムな単語、10%を置き換えない方針で置換する。

2. 隣接文の判定

   質疑応答や自然言語推論といったタスクでは 2 つの文章の関係性を理解できる必要がある。しかしながら、文章間の関係性は、単語の発生確率をモデリングしている言語モデルのみではとらえきれない特徴量を含んでいる。そこで、隣接文判定問題を解かせることで、文章単位での意味表現を獲得する。具体的には、2 つの文章を与え、文章が隣り合っているかを 2 値判定する。2 つの文章の片方を 50%の確率で別の文章に置換する。

BERT モデルは multilingual モデルとして日本語のモデルも提供されている。しかしながら、トークン区切りが日本語に適切なものでなく、これを日本語に適したものに変更する必要がある。英語版の BERT モデルと同様にトークン単位で日本語を取り扱うために、SentencePiece [Kudo 18] を使用する。SentencePiece は、日本語のようなスペースで単語が区切られていない言語にも対応したトークン区切りのアルゴリズムであり、辞書に基づく単語列ではなく、文章から直接トークン区切りを高速に学習する。我々は日本語の Wikipedia をコーパスとして、SentencePiece を使った日本語版の BERT モデルの事前学習を行う。

3. 実証分析

3.1 BERT モデルの事前学習

まずはデータセットとして、Wikipedia 日本語版からコーパスを作成する。コーパスの作成には、2018 年 12 月 20 日時点のダンプデータを用いた。本文の抽出にはオープンソースプロジェクトである WikiExtractor*1 を利用し、不要なマークアップを取り除いた。その後、不要なタグの削除、空白・改行削除、大文字小文字変換などの正規化処理を行い、学習に必要な前処理を施した。BERT モデルの事前学習の際には隣接文の判定のため、Wikipedia の一記事ごとに偶数個の文章からペアを作成し適用する。なお BERT モデルのサイズでは、[Devlin 18] の $\text{BERT}_{\text{BASE}}$ と同様の設定を行った。

3.2 景気ウォッチャー調査を用いたセンチメント分析

センチメント分析のための評価データに内閣府が調査・公表している景気ウォッチャー調査を用いた。景気ウォッチャー調査とは内閣府が行う景気動向に関するアンケート調査で、タクシー運転手や小売店の店主等、景気に敏感な人達 (景気ウォッチャー) に景気に対する判断 (5 段階で、現状と先行きの 2 種類がある) とその理由を毎月アンケート調査している。5 段階の判断はそれぞれ、「良い」、「やや良い」、「変わらない」、「やや悪い」、「悪い」である。これら全件のテキストデータと景気判断が CSV で利用可能となっている。実際の回答サンプルや質問項目の詳細については [内閣] を参照。

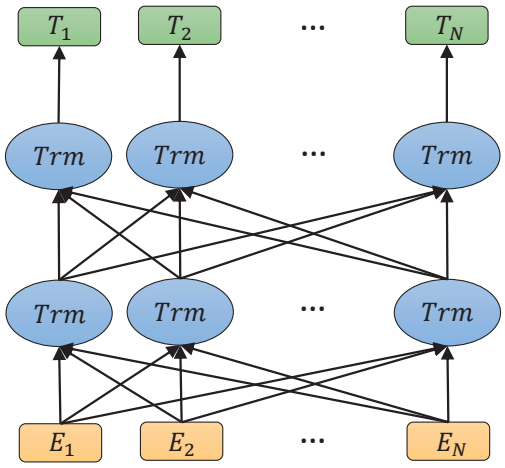

図 1: BERT model の構成。[Devlin 18] より著者作成。Trm は Transformer を表す。

表 1: LSTM を用いたセンチメント分析

	precision	recall	F1 score	support
良い	0.69	0.60	0.64	1,715
やや良い	0.86	0.86	0.86	18,935
変わらない	0.84	0.85	0.84	41,601
やや悪い	0.76	0.77	0.76	23,866
悪い	0.71	0.65	0.68	8,427
Total	0.772	0.746	0.756	94,544

これらのデータのうち、ランダムに 6 割を学習データ、2 割を検証データ、残りの 2 割を評価データとして用いて実験を行う。景気に対する判断のテキストを入力に、そのラベル (5 段階の評価) であるセンチメントを予測するタスクを行い、BERT モデルの有効性を確認する。前節の Wikipedia 日本語版 BERT モデルを事前学習として使用し、ファインチューニングしたモデルと、ベンチマークとして先行研究である [山本 16] で用いられた LSTM モデルを使用する。LSTM モデルの学習では [山本 16] と同様の値を採用し、主なパラメーターに中間層の次元数=250、バッチサイズ=50、ドロップアウト=0.3、エポック数=40 を設定した。結果は表 3.2 および 3.2 の通りである。

BERT モデルを使用することで、「良い」の再現率を除き、**他の全てのラベルおいて適合率・再現率・F 値が改善している**のが見て取れる。また、ラベルごとのサンプル数の多寡に寄らない改善が行われていることから、適度な汎化性能を有していると考えられる。これにより、経済テキストデータによるセンチメント分析でも BERT の事前学習が有効であることが確認できた。

*1 https://github.com/attardi/wikiextractor

表 2: BERT による事前学習を用いたセンチメント分析

	precision	recall	F1 score	support
良い	**0.75**	0.58	**0.65**	1,715
やや良い	**0.88**	**0.90**	**0.89**	18,935
変わらない	**0.87**	**0.86**	**0.86**	41,602
やや悪い	**0.77**	**0.81**	**0.79**	23,866
悪い	**0.76**	**0.67**	**0.71**	8,427
Total	**0.806**	**0.764**	**0.78**	94,545

4. まとめ

本研究では、[Devlin 18] の BERT モデルを Wikipedia 日本語版 [Wikimedia Foundation] と SentencePiece を用いて学習し、その事前モデルを用いて内閣府の景気ウォッチャー指標を学習したセンチメント分析を行った。BERT モデルを使用することで、先行研究である [山本 16] のモデルと比較して、適合率・再現率・F 値の改善を確認した。日本語の経済テキストデータによるセンチメント分析においても、BERT モデルによる事前学習の有効性を確認することができたことから、他の応用事例においても活用が期待できるだろう。

参考文献

[Chen 15] Chen, K.-Y., Liu, S.-H., Chen, B., Wang, H.-M., Jan, E.-E., Hsu, W.-L., and Chen, H.-H.: Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 8, pp. 1322–1334 (2015)

[Devlin 18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018)

[Kudo 18] Kudo, T.: Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates, *arXiv preprint arXiv:1804.10959* (2018)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)

[Pang 02] Pang, B., Lee, L., and Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86Association for Computational Linguistics (2002)

[Turney 02] Turney, P. D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424Association for Computational Linguistics (2002)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, in *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)

[Wikimedia Foundation] Wikimedia Foundation, I.: Wikipedia 日本語版: `https://ja.wikipedia.org/`

[山本 16] 山本裕樹, 松尾豊 : 景気ウォッチャー調査の深層学習を用いた金融レポートの指数化, in *The 30th Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 1–4 (2016)

[内閣] 内閣府 : 景気ウォッチャー調査 : `http://www5.cao.go.jp/keizai3/watcher/watcher_menu.html`

[和泉 17] 和泉潔, 坂地泰紀, 伊藤友貴, 伊藤諒 : 金融テキストマイニングの最新技術動向, 証券アナリストジャーナル』, 日本証券アナリスト協会, Vol. 55, No. 10, pp. 28–36 (2017)