対話型 AI を活用した投資知識などの推定

Estimation of Financial knowledge by Conversational AI

河合継 [1*]　新田翔 [1]　木村祐輔 [1]　眞嶋啓介 [1,2]　西山昇 [3]
Kei Kawai[1]　Sho Nitta[1]　Yusuke Kimura[1]　Keisuke Majima[1,2]　Noboru Nishiyama[3]

[1] クリスタルメソッド株式会社
[1] Crystal method Co. Ltd.
[2] 慶応義塾大学環境情報学部
[2] Keio University Environment and Information Studies
[3] 千葉商科大学会計大学院　客員教授
[3] MBA Program, Graduate School of Accounting & Finance,
Chiba University of Commerce

Abstract: 本研究では、投資信託の販売促進を行う対話 AI の研究を行った。全体の流れは、Seq2Seq を利用した雑談エンジンを通し会話の結果を返す。会話中の言葉の傾向の分析、会話によるリスク許容度の推定、そしてリスク許容度に応じた投資信託をすすめるという流れとなっている。今回の研究では、投資信託を薦める部分について取り組み、有価証券報告書や運用報告書月次レポートを用いて学習をした。学習器については LSTM の階層的 attention モデルの評価が各種方面で評価が高いのを鑑み、今回の研究でも利用した。また同時に学習が軽量でテキスト分類にもよく用いられる FastText も検証した。

1　はじめに

AI を活用することにより、働き型改革などを政府が唱えるなど、人間が働くことなく今までより便利な社会を作ることが重要視されている。また、人に変わる働き手を AI とすることで、少子高齢化社会でも皆が幸せに暮らせる社会を作る事が目標とされている。金融分野も例外ではなく、今まで銀行の窓口に人が立って、販売を促したり、預金勘定など対応していたが、そこに変わる AI なども提案されつつある。また、昨今の AI 化進展とともに今まで人間でしかできないと考えられていた、高度な作業も AI で行う事が出来るようになってきた。まるで人間が話しているかのように電話で美容院の予約を取るような AI も登場してきている。AI に投資信託を売ることができるかという事を命題として、証券アナリストの方々も検討を重ねているようなインターネット記事なども見られるようになってきた。

また、近年著しく発展している AI 技術を金融市場の様々な場面に応用することが期待されている。自然言語処理 (Natural Language Processing, NLP) の分野では、記憶能力を備え、時系列データを処理可能な、ニューラルネットワークの一つである LSTM (Long Short Term Memory) をはじめとする RNN (Recurrent Neural Network) を用いることで文書分類タスクで良好な結果を挙げている。また、FaceBook 社が作っている FastText もその学習時間の短さや軽快さでユーザーを増やしている。大量文書でも学習が行えることが特徴の一つで、Wikipedia 全文のような 100 万ページ以上の文書でさえも学習を行った実績がある。LSTM に関しては、階層的 attention モデ ルを導入することで、文書分類タスクにおいて、SVM (Support Vector Machine) や LSTM を上回るパフォー マンスが達成されている [3]。

これまで、このような AI 技術を用いて、決算短信や有価証券報告書の要約タスク等は行われてきた。これにより、専門知識を有する複数のアナリストが長時間を費やして要約をしていた運用報告書や月次レポートをはじめとする膨大なテキストデータの分析コストが大幅に削減されることが見込まれている。しかしながら株式や投資信託の販売促進を目的とした解析はほとんど行われていない。そこで本研究では、最終的な目的として、販売促進のニーズが高まっている投資信託のデータから、投資信託の販売促進を行う対話 AI の作成を掲げる。今回はそのプロセスの 1 つとして、投資信託の分類を試みる。

本論文では上述の階層的 attention モデルを用いることで、テキストデータから投資信託のパフォーマンス

*連絡先：クリスタルメソッド株式会社
E-mail: kawai@crystal-method.com

推測の検証を行う。

2 分析手法

2.1 FastText を使った文書分類

FastText は単語ベクトルを生成することができる。その単語を数値表現にしたものを分散表現と呼ぶ。分散表現は単語同士の近さを測ったり、演算（足し算・引き算など）を行う事ができる事が特徴となっている。CBOW と呼ばれる文章の中にある単語を前後の単語から推測するモデルや、skip-gram と呼ばれる、ある単語から、その前後に出てくる単語を推測するモデルが内部実装されている。

CBOW　解説
skip-gram 解説
Bag of Tricks for Efficient Text Classification　引用

2.2 階層的 attention モデル

階層的 attention モデルとは、単語レベルでの attention と文レベルでの attention の、二段階の attention を組み合わせたものである。文書分類を行う際、一般的に全ての単語および全ての文が重要であるとは限らない。金融業界の文書だと判定する場合には、金融に関する専門用語がより重要である。また文は単語から成り、文書は文から成り、文書は階層的構造を持っている。二段階の attention を用いてこれらの特徴を捉えることで、文書分類を高い精度で行うことができる。階層的 attention モデルは単語系列に対する Encoder と Attention、そして文系列に対する Encoder と Attention により構成されている。今回それぞれの Encoder は bidirectional GRU (Gated Recurrent Unit) により構成する。階層的 attention モデルの構造は図 1 に示す。階層的 attention モデル では、単語および文ごとに attention を適用し、重要な 単語や文が強調されることで、文書の階層的構造を捉えることができる。

ある文書に L 個の文 s_i $(i = 1, \ldots, L)$、そしてそれぞれの文に T_i 個の単語 w_{it} $(t = 1, \ldots, T)$ が含まれているとする。この時、単語系列に対する Encoder は、以下のように表される。ただし W_e は単語 w_{it} の分散表現を獲得するための埋め込み行列であり、上付き矢印は双方向 GRU の方向を示している。

$$
\begin{aligned}
x_{it} &= W_e w_{it},\ t \in [1, T], && (1)\\
\overrightarrow{h}_{it} &= \overrightarrow{\mathrm{GRU}}(x_{it}),\ t \in [1, T], && (2)\\
\overleftarrow{h}_{it} &= \overleftarrow{\mathrm{GRU}}(x_{it}),\ t \in [T, 1]. && (3)
\end{aligned}
$$

双方向 GRU の結果を結合し、$h_t = [\overrightarrow{h}_{it}, \overleftarrow{h}_{it}]$ を次の単語系列に対する attention に渡す。attention では、以

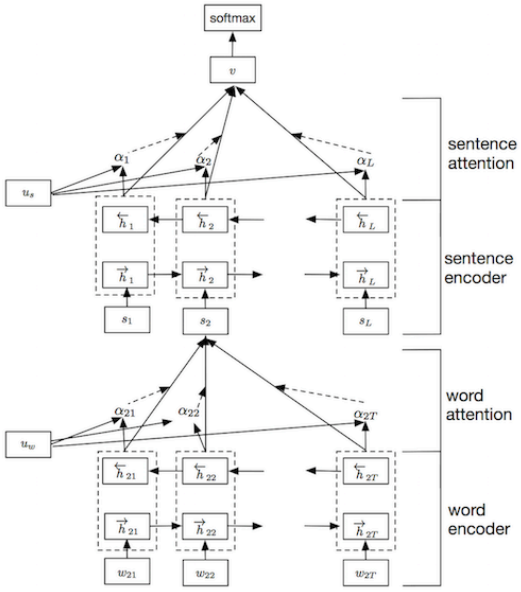

Figure 2: Hierarchical Attention Network.

図 1: 階層的 attention モデル

下に示すように 1 層の全結合層および softmax 関数を用いて、それぞれの単語 w_{it} の重要度 α_{it} を計算する。

$$
\begin{aligned}
u_{it} &= \tanh(W_w h_{it} + b_w) && (4)\\
\alpha_{it} &= \frac{\exp(u_{it}^{\mathrm{T}} u_w)}{\sum_t \exp(u_{it}^{\mathrm{T}} u_w)}, && (5)\\
s_i &= \sum_t \alpha_{it} h_{it}. && (6)
\end{aligned}
$$

次に、文系列に対する Encoder は以下のように表す。

$$
\begin{aligned}
\overrightarrow{h}_i &= \overrightarrow{\mathrm{GRU}}(s_i),\ t \in [1, L], && (7)\\
\overleftarrow{h}_i &= \overleftarrow{\mathrm{GRU}}(s_i),\ t \in [L, 1]. && (8)
\end{aligned}
$$

単語系列のときと同様に、双方向 GRU の結果は結合し、$h_t = [\overrightarrow{h}_i, \overleftarrow{h}_i]$ とし、h_i を文系列の attention に渡す。

$$
\begin{aligned}
u_i &= \tanh(W_s h_i + b_s) && (9)\\
\alpha_i &= \frac{\exp(u_i^{\mathrm{T}} u_s)}{\sum_i \exp(u_i^{\mathrm{T}} u_s)}, && (10)\\
v &= \sum_i \alpha_i h_i. && (11)
\end{aligned}
$$

上式の v は文書に含まれる全文の情報を縮約したベクトルである。文書の分類の際には、v を全結合層に渡し、ソフトマックス関数を利用する。

3　実証分析

3.1　データセット

3.1.1　有価証券報告書のデータセット

有価証券報告書は金融庁が構築している、金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム EDINET と呼ばれるサイトで公開されている。EDINET には年 2 回、5 年分の有価証券報告書が掲載されている。今回の研究では、600 ファンド分、5 年分のデータをスクレイピングにより、EDINET から取得して PDF データ化を行った。各ファンド年に 2 本ずつの PDF2Text ライブラリによりテキストデータを生成し、Quick 社の Qr1 端末から価格データを取得した。有価証券報告書は全文利用して学習データとした。

有価証券報告書データの概要

3.1.2　月次レポートのデータセット

実際の月次レポートおよび価格の時系列データを用いてデータセットの作成を行った。月次レポートに関しては、文章中から運用成績や今後の展望に関わる箇所を抽出し使用した。また正解データに関しては、価格の月次収益率を計算し、その収益率に基づき正解データを作成した。データの概要は表 1 の通りである。正解データは、収益率の中央値を閾値として、閾値に対しての大小に基づいている。

表 1: 月次レポートの概要

データの個数	730
期間	2016 年 1 月〜2019 年 2 月
データ頻度	月次

3.2　分析手法

3.2.1　有価証券報告書の分析手法

有価証券報告書の分析方法としては、当該始値より次の始値が上がっている場合利益（lavel1）、当該始値より次の始値が下がっている場合損益（lavel2）のように文章に分類フラグを付けて FastText の学習関数で学習させた。

ラベルがついている表

2014 年〜2018 年上半期までのデータを学習データとして、2018 年下半期の学習データを評価用データとしてりようして、検証を行った。

3.2.2　月次レポートの分析手法

投資信託の分類として、価格のデータに基づいた、文書の 2 クラス分類を行う。文書データの分散表現獲得には、日本語の Wikipedia のデータから学習させた FastText を使用した。分散表現の次元数は 300 とした。階層的 attention モデルの学習に使用したパラメータは表 2 の通りである。学習エポック数に関しては、学習の進み具合を考慮してある。

表 2: 学習パラメータ

最適化関数	Adam
損失関数	交差エントロピー
学習率の初期値	0.001
学習エポック数	54
バッチサイズ	32
GRU のユニット数	150

3.3　分析結果

3.3.1　FastText

3.3.2　階層的 attention モデル

階層的 attention モデルによる投資信託のリターンの推定についてまとめる。図 2 より、Loss に関しては学習データに対しては 0.35 付近まで下がっているが、検証データに関しては 0.60 程度で下がらなくなっている。Accuracy の関しては、学習データに対しては 80% を超え、検証データに関しても 70% を超えている。

4　まとめと今後の課題

本研究では投資信託の販売促進を目的とした対話 AI の作成を最終目標とし、その取り掛かりとして投資信託のテキストデータから、パフォーマンスの推測を行った。階層的 attention モデルは自然言語処理の分野で高い精度を誇り、投資信託の月次レポートに含まれる運用成績や今後の展望に関する文章から、投資信託のパフォーマンス推測が行えることが期待された。

実際に、階層的 attention モデルによる予測は検証データに対しても 70% を超え、階層的 attention モデルを用いることで、文書分類を高い精度で行えることがわかった。しかしながらデータセットの少なさ等、改善点は残っている。

今後の予定として、投資信託の文書分類の精度向上を目指す。その後は顧客の属性と文書分類を紐づけて、投資信託の販売促進を行う対話 AI の作成に取り組む。

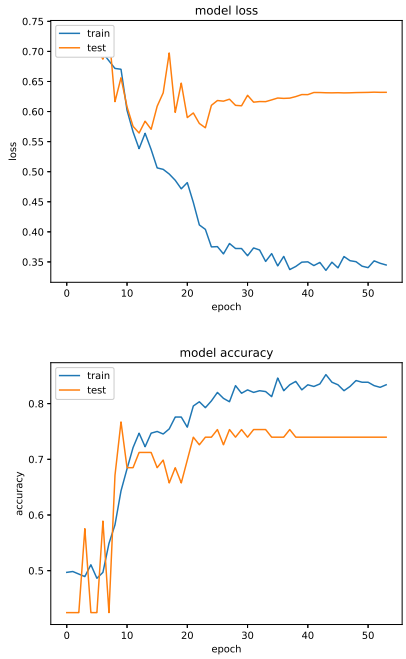

図 2: 階層的 attention モデルの Loss（上）と Accuracy（下）の推移

謝辞

本研究を進めるにあたり、データ提供をしていただいた株式会社 QUICK 様には大変お世話になりました。深く感謝致します。

参考文献

[1] Xiang Zhang, Junbo Zhao, and Yann LeCun: Character-level convolutional networks for text classification, arXiv:1509.01626 (2015)

[2] Duyu Tang, Bing Qin, and Ting Liu: Document modeling with gated recurrent neural network for sentiment classification, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422-1432 (2015)

[3] Yang, Zichao., et al: hierarchical attention networks for document classification, Proceedings of the 2016 Conference on the North American Chapter of the Association for Computational Linguistics, Human Language Technologies. (2016)