# 大規模ニューステキストを用いたナレッジグラフの構築

## Knowledge Graph construction using large-scale news text

張 迎[1]　菅 愛子[1]　高橋 大志[1]

Ying Zhang[1], Aiko Suge[1], Hiroshi Takahashi[1]

[1] 慶應義塾大学大学院経営管理研究科
[1]Graduate School of Business Administration, Keio University

**Abstract:** Investors refer to a variety of information when making investment decisions: The amount of available information is increasing day by day due to the spread of the Internet, and it is difficult to grasp all information such as causality. In this study, we attempt to construct an understanding support model that can visualize news through Attention-Based Bi-LSTM model using Reuters News as an analysis target.

### １．はじめに

投資家は投資判断を行う際，様々な情報を参考にしている，インターネットの普及等により，利用可能な情報量は日々増加しており，因果関係等すべての情報を把握することは困難である．本研究では，ニュースに記載される事象を可視化する理解支援モデルの構築方法を検討し，モデルによる結果を使って資産運用分野における利用の効果を検証した．

### ２．先行研究

因果関係を構築するには,ニューステキストにおけるエンティティ（地名，国家，人名など）の抽出及びエンティティ間のリレーションの抽出という二つの手法を必要とする．エンティティ抽出とリレーション抽出の手法は 2 種類ある．従来は予め辞書を用意して用いていたが，近年は多く自然言語処理技術を使う手法が多く用いられる．本研究では後者に基づき，研究を行う．

エンティティ抽出に関し，本研究で利用する Bi-LSTM-CRF モデルは Lample[1]らにより構築され, エンティティ抽出において LSTM モデル，Bi-LSTM モデル[4]と比較して，より高い精度を示している．

一方，リレーション抽出に関して，本研究では Zhou[3]らが構築した Attention-Based Bi-LSTM モデルを利用する. 当モデルは Bi-LSTM（双方向 LSTM）の上 Attention 層を追加することで，文章を使ったリレーション抽出にて高い精度を示すことが報告されている．

### ３．目的

本研究では，ニュース記事の理解に役立つような背景知識を取得し，可視化する手法を提案する．背景知識の取得には，百科事典記事に記載される構造化データを使わず，非構造データであるニューステキストを用いる．

### ４．データ

単語ベクトルについては, 単語のベクトル表現を取得するための教師なし学習アルゴリズム Glove で Wikipedia を訓練さした単語ベクトルを利用する．

ニュースデータについては, 2010 年 1 月から 2019 年 3 月までの英語のロイターニュースの発信日時とニュースのタイトルを用いる．なお，ニュースデータは予めエンティティやリレーションをつけていないため，エンティティ抽出に使う教師データとして CoNLL-2003 を利用し，リレーション抽出に使う教師データとして Wiki80 を用いる. 検証モデルの構築には，トヨタ自動車(株)の 2015 年 1 月から 2017 年 12 月の株式価格を使用する．

### ５．分析手法

本研究では，先行研究に提示されたモデルを利用し，エンティティ抽出とリレーション抽出を行い，抽出結果を Neo4j により可視化した(Figure 1).

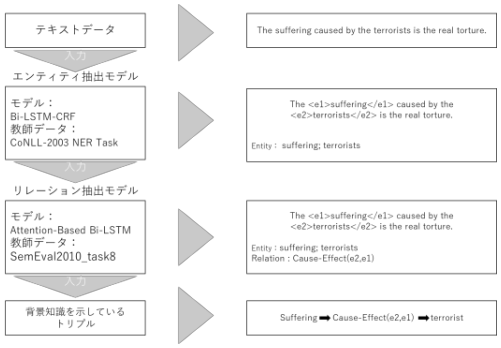


Figure 1. ナレッジグラフの構築手順

資産運用分野における利用効果の検証方法として，リレーション抽出モデルで予測できたトリプルで，菅[5]らが使用した手法のもと，ニュース配信による株式市場の変動の分析を 1 日ごとに分析する．

まず，エンティティ抽出用の Bi-LSTM-CRF モデルとリレーション抽出用の Attention-Based Bi-LSTM モデルにそれぞれ conll2003_en と Wiki80 を用い，モデルの訓練を行う．

訓練されたエンティティ抽出モデルに対し，ニュースデータを入れて事象抽出する．さらに事象の抽出が完了したデータを訓練されたリレーション抽出モデルに入力し，事象と事象の間にリレーションを予測する．結果となる主語，述語，目的語の 3 つの組（トリプル：Triple）を Neo4j でネットワーク図を構成する．

トリプルの検証は，LSTM（Long short-term memory）でニュース記事分析モデルを構築する．モデル化は，トヨタ自動車(株)を例に検証モデルを構築する．具体的に,「トヨタ」というワードが含まれるニュース，及びリレーション抽出の結果により関連性を持つと予測した会社名，人名などが含まれるエンティティのニュースの発信日をトヨタ自動車(株)の株式収益率（閾値 0.1%）により 3 分類（Positive、Neutral、Negative）して教師データとして与える．なお，Neutral のサンプル数が少ないため，本研究では除外した．　教師データを１）業名が含まれるニュース．２）企業に関連しているエンティティのニュース．３）企業名と関連するエンティティ両方のニュース．の三種類とする．

## ６．分析結果

Figure 2 はトヨタ自動車(株)を例にニュースにより作成したトリプルの一部を Neo4j で示したものである．Table 1 は，トヨタに関連を持つエンティティを表に整理したものである．


Figure 2. トヨタに関連を持つエンティティ

Table 1. トヨタに関連を持つエンティティ(表)

| | | | | |
|---|---|---|---|---|
| nissan | bmw | toyoda | sony | avalon |
| kenya | nhtsa | toyota motor corp | ford | edmunds |
| honda | takata | honda motor co | aichi prefecture | karnataka |
| softbank | autodata | panasonic corp | prius | |

リレーション予測モデルを構築する際に使用した訓練データ Wiki80 はビジネスニュースでの適用性がないため，エンティティの間に関係があるかどうかを予測できるが，リレーションの種類はうまく予測できなくて，結果の可読性が低い．

ただし，種類と関係なく，リレーションの自体には有意性があるため，本研究では資産運用分野における利用効果について検証モデルを構築した．

Table 1 に基づいて作成した教師データの中身は Table 2 で表す．Table 3 は 3 つのデータでそれぞれ構築した検証モデルの予測精度である．

Table 2. 検証モデル構築で使用した教師データ

| | Positive | Negative | 合計 |
|---|---|---|---|
| TOYOTA | 1,750 | 1,794 | 3,544 |
| 関連ニュース | 23,541 | 23,030 | 46,571 |
| TOYOTA・関連ニュース | 24,437 | 23,955 | 48,392 |

Table 3. 検証モデルの予測精度

| | 訓練データ | 検証データ |
|---|---|---|
| TOYOTA | 0.7812 | 0.6938 |
| 関連ニュース | 0.8281 | 0.6543 |
| TOYOTA・関連ニュース | 0.7969 | 0.6545 |

検証結果により，トリプルに通じて繋いだエンティティが含まれるニュースは実際に企業の株式価格と関係を持つことを示し，ナレッジグラフは資産運用での有用性を示した．

## ７．まとめと今後の課題

本研究は，ニュース記事の理解に役立つよう，ニュースの背景知識を取得して可視化する手法にいて検討を行った．可視化した結果から特定の企業実体を選出し，それに関連する実体のニュースを当該企業の株式価格の関連性についてモデルを構築して検証した．結果では関連する実体のニュースは企業の株式価格への影響が存在すると示し，本研究の研究手法の有用性を示した．今後の課題は，精度向上のための教師データの選択とモデルのパラメータの調整である．

## 参考文献

[ 1 ] Guillaume Lample, Miguel Ballesteros, Sandeep

Subramanian, Kazuya Kawakami, Chris Dyer: Neural Architectures for Named Entity Recognition, (2016)

[２] J. Lafferty, A. McCallum, and F.C. Pereira: “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proc. 18th International Conference on Machine Learning 2001 (ICML 2001), pp.282–289, (2001)

[３] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi , Bingchen Li, Hongwei Hao, Bo Xu: Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification, the 54th Annual Meeting of the Association for Computational Linguistics, pages 207–212, (2016)

[４] Thireou, T.; Reczko, M.: Bidirectional Long Short-Term Memory Networks for Predicting the Subcellular Localization of Eukaryotic Proteins, IEEE/ACM Transactions on Computational Biology and Bioinformatics 4 (3): 441–446, (2007)

[５] 菅 愛子，高橋 大志：高頻度データを通じたニュースと株式市場の関連性の分析，日本証券アナリスト協会，証券アナリストジャーナル(2018)