6T-08

大規模ニュースステキストを用いたナレッジネットワーク構築*

張 迎† 菅 愛子‡ 高橋 大志‡
慶應義塾大学† 慶應義塾大学‡ 慶應義塾大学‡

１．はじめに

投資において，情報の効率的な獲得は投資における重要なプロセスの一つに挙げられる．近年，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，因果関係等も含めた情報の把握の重要度は高まっている．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た.

２．関連研究

文章の複数の要素間の関係性に着目した研究は数多く報告されている．文章中における因果関係を構築するには，ニュースステキストにおけるエンティティ（事象，地名，国家，人名）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ていれる．本研究では後者に基づき,分析を行う.

エンティティ抽出（Named Entity Recognition）に関し，本研究で利用する Bi-LSTM-CRF モデルは Lample et al. [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[7]，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（Relation Extraction）に関して，本研究では Zhou et al. [3]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．研究の目的

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， 2011 年 1 月から 2019 年 4 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本文を用いる．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 conll2003_en 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.

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．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.

６．分析結果

Figure 1 はホンダ(株)を例にリレーション抽

*Knowledge network construction using large-scale news text
†Ying Zhang, Keio University
‡Aiko Suge, Keio University
‡Hiroshi Takahashi, Keio University

出モデルでエンティティ情報を基に複数の要素間の関連性について特定を行った結果の一部をNeo4j で示したものである.

Figure 1 ナレッジネットワークを可視化した結果

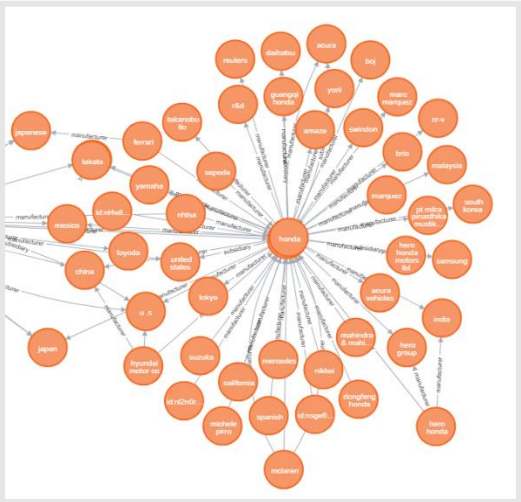

Table 1 はホンダ(株)に関連するエンティティを整理した結果である.

Table 1 ホンダ(株)に関連するエンティティ

支社など関連が強い会社				
hero honda	hero group	guangqi honda	daihatsu	yorii
swindon				
他社				
toyota	boj	takata	yamaha	samsung
製品シリーズ				
acura	cr-v	acura vehicles		
国				
japan	china	mexico	india	u .s
united states	tokyo	south korea		
人物				
takanobu ito	marquez			

ホンダ(株)に関連しているエンティティは主にホンダの製品シリーズ，支社など強い関連を持つ会社，国，人物などがある.

ホンダは自動車安全部品メーカーのタカタ(Takata)と繋がりがあることや，サムスン（Samsung)はホンダと Bluetooth の部品について取引関係を持っていること，マルク・マルケス（Marquez）はロードレースでホンダチームに所属していることなどの関係もモデルに通じて表示することが可能となる.

６．まとめおよび今後の課題

本稿では，ニュースの背景知識を取得して可視化する理解支援モデルの構築を行った．既存モデルの基本性およびテキストデータの確認を行っているが，ナレッジネットワークの実用性の探索，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は今後の課題である.

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] SUN Zhen, W. H.-l. (n.d.). Overview on the advance of the research on named entity recognition[J]. New Technology of Library and Information Service, 2010(6): 42–47.

[６] LIU Xiao-hua, ZHANG Shao-dian, WEI Fu-ru, et al. Recognizing named entities in tweets[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies. Stroudsburg, PA:ACL, 2011:359-367.

[７] 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)