経営課題にAIを!ビジネスインフォマティクス研究会 (第13回)
JSAI Special Interest Group on Business Informatics (SIG-BI #13)

大規模ニューステキストを用いた因果関係ネットワーク構築の試み

Causal network 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,ロイターニュースを分析対象として，Attention-Based Bi-LSTM モデル[3]を通じ，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を試みる．

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（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, 2017 年 1 月から 2018 年 12 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 NER Task を利用し，リレーション抽出に使う教師データとして SemEval2010_task8 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 CoNLL-2003 NER Task と SemEval2010_task8 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（Figure 1）．

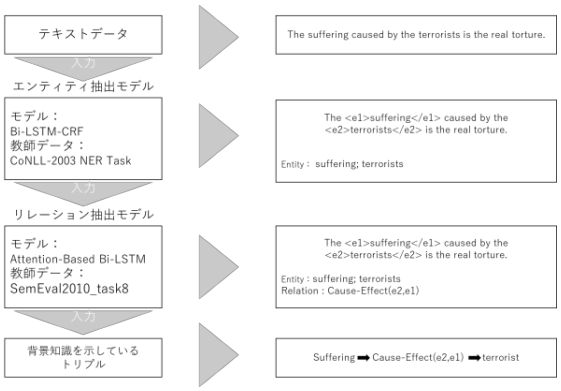

Figure 1.研究手順

６．分析結果及び今後の課題

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．これまで，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，より大規模なニューステキストの使用，及び精度向上のための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)