因果判定データセットの構築と原因結果表現抽出への拡張

仁木 裕太　坂地 泰紀　松島 裕康　和泉 潔
東京大学　大学院工学系研究科
{b2018yniki, sakaji, matsushima, k.izumi}@socsim.org

1 はじめに

近年, Web ページや新聞記事など, 計算機が処理可能な電子媒体の文書データが豊富に存在する. それと同時に, テキストマイニングの手法によって, 価値のある情報を自動で抽出することが可能になりつつある. 文書中に多種多様な情報がある中で, 本研究では「因果関係知識」に注目する. 経済領域において因果関係知識は, 市場動向の分析や機会損失の防止のために有益な情報であるからである. 例えば, Sakaji ら [1] は, 日本企業の財務諸表から出現頻度が小さい因果関係知識を発見する手法を提案している. 彼らは, 周知されていない因果関係知識を抽出することで, 投資信託の機会を発見することを目的としていた.

因果知識の抽出は自然言語処理の重要なタスクで, 多くの研究が行われてきた. Girju [2] は, 手がかり表現を元に英語の文書から因果関係 (名詞句のペア) を自動で検知し抽出する手法を提案した. Khoo ら [3] は, 人手で作成したパターンをもとに, 英語の新聞記事から原因・結果の情報を抽出する手法を提案した. これらの研究はそれぞれ, 「Why 型」の質問応答と医療文書の解析に利用されている. 我々の知る限りでは, どの研究も 1 つの言語の文書を対象に行われており, 複数言語の文書を対象にした研究はされていない.

各言語に対してモデルを作るには多大な労力を必要とし, 複数のモデルをメンテナンスするためには, モデルの数だけコストが大きくなる. 当然, 複数の言語の文章に利用可能なモデルを作ることができれば, より多くの因果関係知識をより簡単に収集することが可能になる. そこで, 本研究では複数言語の文書に利用可能なモデルを構築することに着目した.

多くの研究では, 日本語の場合は「で」や「から」, 英語の場合は “because,” “since” などの手がかり表現を使って, 因果関係知識を抽出している. しかし, これらの手がかり表現は, 例えば手段や起点などの別の意味で使われることがある. また, 文中に因果関係が存在しても, 手がかり表現がないまたは考慮されていない表現が使用されている場合も考えられる. したがって, 文中に因果関係知識が含まれているかを判別するモデルを, 手がかり表現を外生的にに与えずに構築する必要がある.

以上の理由から, 複数言語の文書に使用可能な, 文中における因果関係知識の有無を判別するモデルを本研究で提案する. 具体的には, Long Short-Term Memory (LSTM) というニューラルネットのモデルをベースに, 注意機構を組み込んだモデルを用いた. 日経新聞の記事から作成したデータセットと Semeval という英語のオープンデータセットを対象に, Random Forest (RF) や Logistic Regression (LR) などの機械学習モデルを比較手法として実験を行った.

2 構築するデータセット

2.1 本研究で扱う因果関係

ここでは, 本研究で扱う因果関係について説明する. 我々は因果関係を, 原因となる出来事とその結果発生した事象のペアと考える. 具体例を表 1 に示す. 1 つ目の例では, 「雪」が原因にあたり, 「空の便が乱れ」が結果にあたる. 3 つ目と 4 つ目の例では, 「で、」という手がかり表現が存在するが, これらは前後の情報を並列しているだけで, 因果関係を示すものではない.

2.2 データの説明

本研究では, 坂地ら [4] が日経新聞記事から構築した文中の因果有無判定データセットの仕様にならって, SemEval2010 task8[1]のデータセットを再構築した. それぞれのデータセットのサンプル数を表 3 に示す.

[1] http://semeval2.fbk.eu/semeval2.php?location=data

表 1: 日経新聞データセットのサンプル.

因果関係の有無	本文
有り	雪の影響で、北海道や東北、日本海側を中心に空の便が乱れ、十四日午後一時までに計百七十便が欠航、約二万人に影響が出た
有り	販売・生産の低迷脱却が遅れることから、連結最終損益の黒字転換時期も一期ずれ込んで二〇〇七年三月期になる見通し
無し	国際電話でも昼夜の区別のない二十四時間均一料金にしたうえで、大幅な値下げに踏み切る下地が整う
無し	二年連続の前年実績割れで、過去十年間で最低水準となる

坂地らの日経新聞データセットは, 1990 年から 2005 年の記事を無作為に抽出した後, 文中に原因-結果情報が含まれているかを 5 人でタグ付けして作成された. 5 人中 3 人以上のタグが一致したものを学習データとして採用した.

SemEval 2010 task8 は Shared task として公開されているデータセットで, 文中の 2 単語の関係を分類する 9 クラス分類のタスクである. その中に原因–結果の関係もあるため, 本研究では原因–結果の関係かそれ以外かという 2 値分類の形式にデータセットを再構築した.

それと同時に, 原因–結果抽出タスクに利用できるように, SemEval について再度アノテーションの作業を行っている. 例えば, "The <e1 >injury </e1 >resulted in numerous <e2 >operations </e2 >to save his eyesight. "というデータを, "The injury <c >resulted in </c ><r >numerous operations to save his eyesight </r >. "(: 原因, <r >: 結果, <c >: 手がかり表現) というようにタグを付け替えている. 経済事象の分析には, 単語レベルの関係では不十分で, 原因・結果表現が必要だからである. 本再アノテーション作業中に, 以下の 2 つの問題を発見した.

1 つ目は, 我々の目的とは違った因果関係にタグが付けられていることである. 例えば, "The <e1 >microphone</e1 >converts sounds into an electrical <e2 >signal</e2 >. "は, 原因–結果のラベルが付けられていたが, これは道具とその産物の関係で因果関係ではない. 本論文を執筆時点で, train データの 1003 個のサンプルの内 890 個のサンプルについて再アノテーションが完了しているが, 完了している 890 個のうち 67 個のサンプルについて同様の問題が確認された.

2 つ目は, 単語間の関係が原因–結果でなくても, 文中に因果関係が含まれる場合があることである. 例えば, "because"は因果関係の明らかな手がかり表現であるが, 本来の原因–結果ラベルが付いたサンプルの文中では 9 回出現しているが, 他のラベルの文中では 46 回出現している. "therefore, " や"thus" など他の手がかり表現についても, 同様の問題が見られたので表 2 に示す.

表 2: 明確な手がかり表現の出現頻度.

手がかり表現	原因–結果ラベルサンプル中の出現頻度	他ラベルのサンプル中の出現頻度
therefore	1	8
thus	4	16
hence	0	2

以上の 2 つの問題が確認されているため, SemEval のデータセットを原因結果表現抽出に対応するように再アノテーションを行うことが必要であることが言える.

3 因果有無判別モデル

現在, 因果関係の有無についてのデータセットは既に構築が完了したので, 以下のモデルを用いて簡単に実験を行う.

Recurrent neural networks (RNNs) は, 様々な自然言語処理タスクで優れた性能を発揮したニューラルネットのモデルである. 一方, RNNs では勾配消失や勾配発散の問題が発生すると報告されている. 本研究では, これらの問題を解決する機構が組み込まれた Long Short-Term Memory (LSTM) [5] を用いた. 加えて, 言語には前方向と後ろ方向の係り受け関係があるため, その両方を考慮できるよう双方向 LSTM を使用した.

表 3: 本研究で用いたデータセットのサイズ.

データセット	全データのサンプル数	正例のサンプル数	学習用のサンプル数
日経新聞	2001	855	1500
SemEval	10717	1331	8000

また, 近年, 多くの自然言語処理の研究で注意機構による精度の改善が報告されている. これまでタスクに応じて多様な注意機構が提案されているが, 本研究では, 文分類のタスクについて提案された Lin らの注意機構 [6] を用いた.

4 実験

本論文では, 英語と日本語の両方のテキストデータに使用可能な, 文中に因果情報が存在するかを分類する学習モデルを提案する. 実験を通して, その有効性を示す。第 3 節で説明した LSTM および注意機構付き LSTM は, PyTorch で実装した. 加えて, RF と LR, Support Vector Machine (SVM) の機械学習モデルをベースラインとして実験を行った. これらの機械学習モデルは, scikit-learn[2]を用いて実装した.

また, 前処理として分かち書きを行うために 2 つのツールを用いた. 1 つ目は, MeCab[3]と Neologd [7] で, それぞれ形態素解析エンジンとその辞書の役割を果たす. 日本語のテキストデータにこれを用いた. 2 つ目は, spaCy[4]というツールで, 英語のテキストデータに利用した. トークン化の処理と同時に, lemmatization の処理も spaCy で行った.

LSTM の入力は, 文中の単語に対応する id 列である. 単語埋め込みは end-to-end で学習する場合と word2vec (CBOW) を使用する場合の 2 通りで実験した. 日本語の word2vec は Wikipedia で学習したモデル, 英語は Google が提供しているモデル [8] を使用した. ベースラインの機械学習モデルの入力は, ユニグラムから n-グラムまでの出現頻度とした. この時, n-グラムの n と特徴量として考慮する最低出現頻度はパラメータサーチで探索した.

各モデルの探索対象としたハイパーパラメータを表 4 に示す. グリッドサーチでパラメータ探索を行っており, f1 スコアが最も良いハイパーパラメータのセットを採用した.

[2]scikit-learn: http://scikit-learn.org/stable/
[3]https://taku910.github.io/mecab/
[4]https://spacy.io/

4.1 結果と考察

各モデルの性能を表 5 に示す. 性能の評価指標には, Macro f1 と AUROC を用いた.

どちらのデータセットでも, 上位 2 つのモデルが word2vec を用いた LSTM のモデルとなった. このことから, 別のコーパスで単語埋め込みの事前学習モデルが, 精度改善に貢献することがわかる. また, 注意機構付きの LSTM が end-to-end でもトップ 2 のモデルに近い性能を示している. これは, end-to-end でも精度の良いモデルを訓練できる可能性を示唆している.

一方で, 日経新聞データセットに対する LR の結果を見ると, word2vec を用いた LSTM に次ぐ精度を示している. このことから, 日経新聞データセットを使った本研究での問題設定は比較的容易であることが考えられる.

5 結論と今後の展望

本研究では, 文中の因果関係有無を判定するための日本語と英語のデータセットを構築し, LSTM 及び注意機構を用いたモデルで実験を行った. 併せて, 現在 SemEval のデータセットについて作業中の, 原因結果抽出のためのデータセットへの再アノテーション作業が完了すれば, 公開することを考えている. 今後, 経済文書から同じ目的のデータセットを構築することを考えている.

謝辞

本研究は, 大和証券投資信託委託株式会社との共同研究の成果物である.

参考文献

[1] Hiroki Sakaji, Risa Murono, Hiroyuki Sakai, Jason Bennett, and Kiyoshi Izumi. Discovery of rare causal knowledge from financial statement

表 4: 探索したハイパーパラメータ一覧

モデル	ハイパーパラメータ
Random Forest	n, 最小出現頻度, # of estimators, max depth
Logistic Regression	n, 最小出現頻度, C, penalty
Support Vector Machine	n, 最小出現頻度, C, gamma
LSTM	エポック数, バッチサイズ, { 埋め込み, 隠れ層, 中間層 } の次元, 学習率, Dropout 率
LSTM with attention	エポック数, バッチサイズ, { 埋め込み, 隠れ層, 中間層 } の次元, 学習率, Dropout 率

表 5: 各モデルの性能を示した表. トップ 1・2 の精度が太字で書かれている.

model	embedding	日経新聞		SemEval	
		f1	AUC	f1	AUC
Random Forest	1-n gram	0.743	0.788	0.528	0.683
Logistic Regression	1-n gram	0.833	0.854	0.819	0.878
Support Vector Machine	1-n gram	0.830	0.852	0.779	0.832
LSTM	end-to-end	0.806	0.830	0.827	0.888
	word2vec	**0.838**	**0.858**	**0.856**	**0.922**
LSTM with attention	end-to-end	0.831	0.853	0.847	0.902
	word2vec	**0.841**	**0.861**	**0.869**	**0.930**

summaries. In *The 2017 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics*, pp. 602–608, 2017.

[2] Roxana Girju. Automatic detection of causal relations for question answering. In *In ACL Workshop on Multilingual Summarization and Question Answering*, pp. 76–83, 2003.

[3] Christopher S.G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung Hyon Myaeng. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, Vol. 13, No. 4, pp. 177–186, 1998.

[4] 坂地泰紀, 増山繁. 新聞記事からの因果関係を含む文の抽出手法. 電子情報通信学会論文誌 D, Vol. 94, No. 8, pp. 1496–1506, 2011.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[6] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *ICLR*, 2017.

[7] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.