注釈者のクラスタリングに基づく議会発言の事実検証可能性の推定

白土 大樹, 秋葉 友良, 増山 繁
豊橋技術科学大学

d173327@edu.tut.ac.jp

1 はじめに

近年, スマートフォンやインターネット, SNS が世界中で普及し, これにより, いつどこにいても, だれであっても簡単に情報を世界に向けて発信することが可能になった．しかしながら, それに伴い, 正しい情報のみからなるニュースだけでなく, 誤った情報，不正確な情報, 捏造された情報を含むニュースなどもまた広く社会に拡散されている．これらはフェイクニュースと呼ばれ, 社会に深刻な影響を与えている [1] ．実際に 2016 年にはアメリカで, 架空の事件「ピザゲート事件」が起こったというフェイクニュースが拡散され, これを信じた人による発砲事件も発生している．

フェイクニュースの対策としては, SNS などのユーザーが情報を受け取るときに常にその情報が事実かどうか検証することが考えられる．しかし, 日々膨大な情報が国境をも超えて拡散されている現代において, ユーザーが受け取る情報の全てについて事実検証を個人で行うことは困難である．そのため, ユーザーに代わって事実検証を行う組織も存在する．フランスでは, 選挙の時期を狙って拡散されるフェイクニュースに対して事実の検証を行う CrossCheck という Web サイトが立ち上げられた [2] ．これにはその後さまざまなメディアが協力を行った結果, フランス大統領選挙に合わせて拡散されたいくつかのフェイクニュースを暴くことに成功している．しかし, このような組織の情報網とリソースをもってしても日々世界で拡散される情報の全てについて事実検証を行うことはできていない．

そのため, コンピュータにより情報の事実検証を支援する機能が望まれている．しかし, 社会に拡散される全ての情報が事実検証可能であるとは限らず, 検証がそもそも不可能な情報も存在する．そこで, 本研究では事実検証の前段階である事実検証可能性に注目する．議会会議録のデータセットを用いて, 議会における発言に対して事実検証可能性を推定するシステムを開発することを目指した．

2 データセット

本研究では, NTCIR14 QALab-PoliInfo の Formal-Run 用のデータセットを使用する．これは東京都議会の会議録から作成されたデータセットであり, 14 回分の会議録に, 10,291 発言が存在する．発言に対するラベル付けは 20 人の注釈者が行っており, 1 回分の会議録中の発言に対して 3 人もしくは 5 人の注釈者がラベル付けを行っている．ここで, 第三者が真偽を確認可能な事実を根拠として定義し, 根拠を含んだ発言を事実検証可能な発言とする．データセットに存在する発言の例を表 1 に示す．

表 1: 発言例

子ども医療費助成事業は、９月から中学卒業まで拡充するというものの、対象は子供が３人以上いる世帯に限るという全国ではほとんど例を見ない制度のままとなっています。

この発言に対しては注釈者 D, E, F, G, H の 5 人がラベル付けを行った．注釈者 D, G, H の 3 人は根拠なしとラベル付けを行い, 注釈者 E, F の 2 人は根拠ありとラベル付けを行った．このように, 同じ発言であってもそれに対するラベル付けの結果は注釈者ごとに異なっている場合がある．

注釈者間でのラベル付け結果の一致率を調べた結果, 一致率は高い場合は 0.941, 低い場合は 0.213 と, ばらつきが大きく, 著しく低い場合があることが分かった．よって, 注釈者ごとにラベル付けの基準は大きく異なっていると考えられる．そのため, 全ての注釈者のラベル付け基準に合うように分類器を構築するのは難しい．一方，ラベル付け基準を共有する注釈者グループが特

定できれば，その基準に基づく分類器を構築することができると考えられる．

3　注釈者クラスタリング

3.1　クラスタリング方法

ラベル付け基準が近い発言データを得るために，注釈者間でのラベル付け結果の一致率をもとに注釈者のクラスタリングを行う．これによりラベル付けの一致率が高く，ラベル付けの基準が近い注釈者クラスタを得ることができ，その注釈者クラスタがラベル付けした発言データを訓練に用いることで，その注釈者クラスタのラベル付け基準に合わせた分類器の構築が可能になると考えられる．注釈者をノード，注釈者間のラベル付けの一致率を重み付きエッジとする注釈者グラフを作成し，グラフベースのクラスタリングを行う．クラスタリングのアルゴリズムとしては Newman アルゴリズムを用いる [3]．Newman アルゴリズムではグラフのクラスタリング結果の良さを示す指標である Modularity を定義し，Modularity を最大化することを目指してボトムアップでクラスタリングを行う．Modularity は式 1 のように定義される．

$$Q = \frac{1}{2M}\sum_{vw}(A_{vw} - \frac{k_v k_w}{2M})\delta(C_v, C_w) \qquad (1)$$

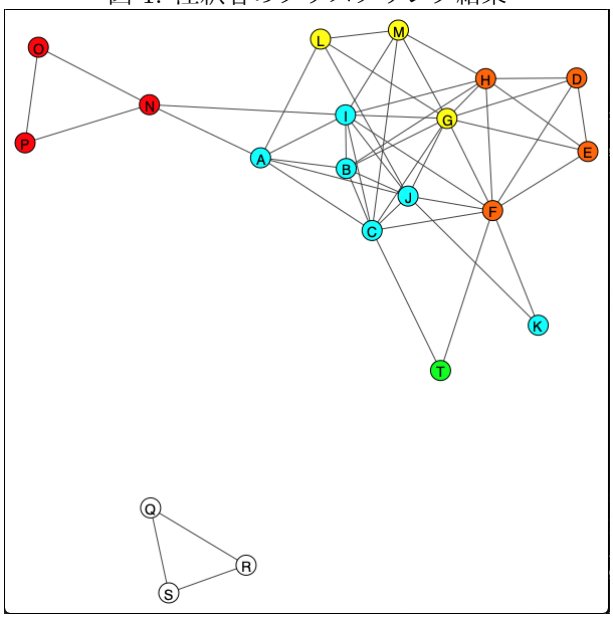
ここで，M はグラフに存在するエッジの総数である．A はグラフの隣接行列であり，A_{vw} はグラフの隣接行列の vw 要素を表す．k_v, k_w はノード v，ノード w の次数である．δ はクロネッカーのデルタであり，ノード v，ノード w が属するクラスタ C_v，クラスタ C_w が同一クラスタである場合には 1，それ以外の場合には 0 になる．ただし，式 1 の定義はエッジの重みを考慮していない．エッジの重みを考慮すると Modularity は式 2 のように定義し直される [4]．

$$Q = \frac{1}{2W}\sum_{vw}(A_{vw} - \frac{w_v w_w}{2W})\delta(C_v, C_w) \qquad (2)$$

W はグラフに存在するエッジの重みの総和であり，w_v, w_w はノード v，ノード w に接続されているエッジの重みの総和である．Newman アルゴリズムによるクラスタリングは以下の手順で行われる．

1. 全てのノードを別々のクラスタに割り当てる．
2. 2 つのクラスタをマージした場合の Modularity を全てのクラスタの組み合わせについて求める．
3. 最も Modularity が高くなる組合せでクラスタをマージする．
4. 手順 2, 3 をクラスタのマージができなくなるまで繰り返す．
5. 最も Modularity が高かったマージ結果をクラスタリング結果とする．

図 1: 注釈者のクラスタリング結果



3.2　クラスタリング結果

注釈者のクラスタリングを行った結果を図 1 に示す．6 つの注釈者クラスタが得られたが，このうち 注釈者 T のみからなるクラスタを除外し，残りの 5 つの注釈者クラスタでクラスタ内の注釈者間でのラベル付け一致率の平均を求めた．結果を表 2 に示す．

表 2: 平均ラベル付け一致率

注釈者	ラベル付け一致率
L, M, G	0.921
H, D, E, F	0.782
A, B, C, I, J, K	0.573
Q, S, R	0.461
O, P, N	0.445

表 2 より，注釈者 L, M, G からなるクラスタが最もラベル付けの一致率が高い．そこで，このクラスタの

表 3: 発言データ例

次に、三番目の大きな項目、都市、集落のあり方についてから、まず、大都市制度と中京都構想についてお聞きをしてまいります。
この地域の社会・経済活動を支える総合交通体系の整備充実は、中京都構想の推進にとっても極めて重要であります。
また、知事会におきましても、子ども医療費助成制度を創設するように提言をしております。

表 4: 発言ではないデータ例

｜　　｜（２）道州制の導入について　　｜吉川副知事　　｜
｜請願第 57 号　　｜正規の通級指導担当教員の増配置及び特別支援非常勤講｜採　　択｜　　要　　｜
｜ 18 ｜米海兵隊ＭＶオスプレイの配備及び飛行訓練の拒否を　｜　｜　｜○｜（多数をもって決定）｜

注釈基準のモデル化を試みた．評価用データとして，注釈者 L, M, G がラベル付けした会議録の 1 割である 589 発言を用いる．残り 9 割の 4,022 発言を注釈者 L, M, G のラベル付け基準に合わせた分類器の訓練に用いる．一方，評価データを除く全注釈者の発言データ 9,254 発言で訓練した分類器も構築し，両者の比較を行う．

4　事実検証可能性の推定

4.1　単語辞書の作成

分類器の訓練を行う前に, 分類器の語彙を決定する単語辞書を作成する．Neologd 辞書 [5](2018 年 7 月 31 日に取得) を追加した形態素解析エンジン MeCab[6] を用いて訓練用データの発言を分かち書きし, 出現した単語を語彙として辞書に登録する．

本研究で使用するデータセットは，基本的には東京都議会の議事録から抽出した発言から成るが，一部発言ではない質問事項表などが含まれている．発言であるデータと発言ではないデータの例を，それぞれ表 3, 表 4 に示す．発言ではないデータには, 記号「—」や「○」が含まれている．これらの記号を分類器の単語辞書に含めないようにするため，品詞が記号, 品詞細分類 1 が一般, それ以外の情報がなし，の形態素は除外した．また，単語辞書作成のときに MeCab による形態素解析の結果, ある単語が数を表す単語であった場合, その単語は<NUM>というラベルに置き換えて辞書に登録を行った．

注釈者 L, M, G がラベル付けした訓練用の発言データから単語辞書を作成した結果, <NUM>ラベルを含めて語彙数 13248 の単語辞書となった．分類器による発言の事実検証可能性の推定のとき, 発言に含まれるこの辞書に存在しない単語は未知語として扱う．

4.2　分類器の構築

LSTM(Long short-term memory)[7][8] を使用した再帰型ニューラルネットワークにより発言から事実検証可能性の推定を行う．分類器は LSTM 層と出力層の 2 層で構成される．

分類器の訓練では, 単語辞書の作成と同様にして, 発言を MeCab によって分かち書きし, 単語の系列にする．その後, 単語辞書をもとに単語を one-hot ベクトルに変換し, 埋め込み層を経て LSTM に入力する．LSTM は入力単語ごとに, 200 次元の隠れ状態を得る．最後の単語を入力した後の LSTM の隠れ状態から，全結合層と softmax 層により事実検証可能性の推定を行う．

分類器の訓練ラベルには，事実検証可能か否かを表す 2 値ラベルではなく，事実検証可能と判定した注釈者の割合から求めた確率的ラベルを用いた．表 1 の発言を例にすると, この発言には 5 人の注釈者がラベル付けを行い, 注釈者 E, F の 2 人は根拠があるため事実検証可能な発言とラベル付けを行い, 注釈者 D, G, H の 3 人は根拠がないため事実検証不可能な発言とラベル付けを行っている．よって, この発言は事実検証可能な確率が 0.4, 事実検証不可能な確率が 0.6 の発言として扱う．損失関数には確率ラベルと予測結果の確率分布の差を表す Kullback-Leibler Divergence を使用した [9] ．LSTM の隠れ状態とセルは 0 で初期化し, パラメータの最適化には Adam を使用した [10] ．学習時の損失の収束状況より, 学習率は 0.0002 に設定し, 100 エポック学習を繰り返した．

4.3　事実検証可能性の推定結果

全ての発言データを使用して訓練した分類器と, クラスタリングによって得られた注釈者 L, M, G のラベ

表 5: 分類器の性能比較

学習データ	全注釈者	L, M, G
Precision	0.763	0.802
Recall	0.527	0.898
F1 値	0.622	0.826
Accuracy	0.531	0.737

ル付けした発言データを使用して訓練した分類器で事実検証可能性の推定を行う．ただし, 本実験で構築した分類器が出力するのは発言が事実検証可能な確率と事実検証不可能な確率であるため, 分類器が出力した確率の高いほうをその分類器による発言の分類結果として扱う．2 つの分類器による事実検証可能性の推定結果の比較を表 5 に示す.

全てのデータを使用して訓練した分類器の Accuracy に注目すると 0.531 であり, 分類がほとんど行えていない．よって, ラベル付けの基準が異なるラベル付けデータを使用した場合, 分類が正しく行えないことが確認された．また, 注釈者 L, M, G がラベル付けを行ったデータを使用した分類器は Precision, Recall, F1 値, Accuracy の全てが, 全てのデータを使用して訓練した分類器を上回っている．注釈者のクラスタリングによってラベル付け基準が近いデータを用いることで分類器の構築が可能になることが確認された.

5 おわりに

本研究では会議録中の発言の事実検証可能性を推定することを目標とし, LSTM を使用した再帰型のニューラルネットワークの分類器を構築した．学習データにおいて注釈者間のラベル一致率が低いという問題に対して，注釈者のラベル付け基準が多様であると仮定し，ラベル付け一致率の高い注釈者をクラスタリングすることによって，クラスタ毎の学習データを構築しそのラベル付け基準に基づく分類器の構築を試みた．評価実験により，ラベル付け一致率の最も高い注釈者クラスタについて性能の高い分類器を構築できることを確認した.

今後は，他のクラスタでの分類器の学習と評価を行う予定である．また，今回得たクラスタリングの結果を詳しく分析することにより，事実検証可能性のための良いラベル付け基準の確立を試みたい.

謝辞

本研究は JSPS 科研費 16K00153 の助成を受けた。

参考文献

[1] 吉田大輔 BuzzFeed FoundingEditor. 真偽が危ういフェイクニュース時代の総選挙日本でもファクトチェックが始まった, October 10 2017. https://www.buzzfeed.com/jp/daisukefuruta/fake-vs-fact-check-in-japan.

[2] Crosscheck, 2017. https://crosscheck.firstdraftnews.org/france-fr/.

[3] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E 70, 066111 (2004)*, 2004.

[4] Joerg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E 74 (2006) 016110*, 2006.

[5] Neologism dictionary based on the language resources on the web for mecab-ipadic. https://github.com/neologd/mecab-ipadic-neologd.

[6] Mecab: Yet another part-of-speech and morphological analyzer. http://taku910.github.io/mecab/.

[7] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, nov 1997.

[8] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with lstm. *Neural Comput.*, Vol. 12, No. 10, pp. 2451–2471, October 2000.

[9] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, Vol. 22, , 1951.

[10] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.