文学作品における教師なし話者同定

遠田 哲史
東京大学大学院 情報理工学系研究科
tohda@tkl.iis.u-tokyo.ac.jp

吉永 直樹
東京大学 生産技術研究所
ynaga@iis.u-tokyo.ac.jp

1 はじめに

我々はニュースや小説など複数の登場人物が現れる文章に触れる時，文章内の発話表現を通じて文章の背景となる社会構造などへの理解を深める．具体的には，発話の内容とその話者と聴者，および対話行為やその場面などの情報を把握することで，登場人物の人間関係や社会的な立場，話の筋などを理解する．

文学作品では特に，登場人物が多数かつ複雑に関係し合っているため，発話と話者を関連付けるデータの重要度が高い．このように関連付けられたデータは，文章内の人物が形成しているコミュニティを可視化させる対話ネットワークの構築 [1] や対話における流暢な応答生成 [2] などに用いることができる．したがって，文学作品内の各発話に対し話者を自動的に同定するタスクを解くことで，膨大な数の文学作品の分析や対話生成用のデータを用意することができると言える．

話者同定タスクに対しては，各発話を実体 (entity) としての話者に関連付けるタスクと，その部分問題として，発話の近傍文章に現れる話者への言及 (mention) に関連付けるタスクが試みられてきた．これまで，発話を話者の entity へ関連付ける問題を解く研究では，発話者一覧および各人物の詳細を利用しているため，これらの手法を実世界の文学作品に適応する際には，高コストの前処理が必要となってしまう．

そこで，本稿では文学作品の話者同定を，発話者一覧等の情報を用いないという意味で教師なしで行うタスクを新たに取り組み，これを発話内容とその周辺状況を考慮してクラスタリングで解く手法を提案する．本手法により，各発話は話者の同一性によってグルーピングされる．この際，予めラベル付けされた発話や発話者の性別・別称などの情報は用いることができない．そこで，提案する手法では，ルールベースの手法で mention を抽出した後に，発話文をベクトル表現に変換し，具体的な人名を mention とする発話を手がかりとして距離関数を学習し，クラスタリングを行う．

実験では Emma データセット [3] を用いて提案手法の評価を行った．その結果，提案手法は発話内容のみを手がかりとするベースラインを上回る結果となった．

2 関連研究

話者同定タスクの初期の研究 [4, 5] では，発話を示す動詞（said 等）を利用し，その前後に位置する名詞と発話を関連付けている．しかし，これらの手法では話者が明示されていない発話を関連付けることが出来ない．また，たまたま主語が固有名詞（人物名）である場合を除き，entity への関連付けを行っていない．Almeida らの研究 [6] では，発話からの話者（mention）同定と共参照解析を同時に解くモデルを学習している．

文学作品を扱った研究としては，まず Elson ら [7] の研究が挙げられる．著者らは各発話について話者の候補のリストを構築し，リストから一人選ぶ分類問題として解いた．タスクとしては entity に至る話者同定を行っているが，テスト時に正しい話者ラベルを使用している．O'Keefe ら [8] の研究では話者同定を系列ラベリング問題として解いたが，文学ドメインでの性能はルールベースのベースラインを下回っている．また，CRF を用いて同じタスク設定を解いた Yeung ら [9] 同様，話者同定は mention までであり，entity への関連付けは考慮していない．

本稿で提案するタスクに最も近いものとして，He ら [10] および Muzny ら [3] の研究があり，これらの研究では話者を entity へ関連付けるタスク設定で解いている．Muzny らはルールベースのふるいを 2 つ設け，発話と mention，mention と entity を紐付けている．これに教師あり学習で学習させた分類器を加えたシステムで，コーパス間の平均 F-score が 87.5 であったと報告している．ただし，両研究とも，全ての発話者の性別や別称などの情報を含む外部データを活用している点で本稿で提案するタスクと異なる．

3　教師なし話者同定タスク

本節では，本研究で新たに取り組む教師なし話者同定タスクについて定義する．我々は，文章とそれに含まれる発話を入力とし，発話者一覧等の文書に依存する教師情報や，話者同定タスクに関するラベル付けされた発話データを用いずに，それぞれの発話に対する話者を文章内から抽出してラベル付けするタスクとする．なお，対象とする発話文については，話者の発話内容が引用符の間に書かれている直接引用を対象にする．O'Keefe ら [8] によれば，英語文章からの発話抽出率は，単純なパターンマッチングで 99%以上となるため，本タスクではシステムの入力時点で発話が完全に抽出されている状態を考える．

2 節で述べた既存研究で取り組まれているタスクと本タスクの大きな違いとしては，発話者の情報が事前に与えられていないことが挙げられる．すなわち，与えられた人物の一覧から候補を絞り込むことで発話者を同定するルールベースを用いることは困難であり，話者 entity の表現が異なる発話同士の関連付けも決して自明ではない．

加えて，教師なしのタスク設定では教師ラベルで学習する機械学習的手法の適用が難しく，この点について工夫が必要となる．

データセット　本研究では QuoteLi3 [3] に含まれている小説 Emma の一部分をデータセットとして用いる．この小説は 19 世紀初頭に書かれた英文小説であり，冒頭より 3 章がアノテーションされている．文章内の全ての発話に対し，それぞれの発話者の mention と mention が指し示す entity が xml 形式で記述されている．Mention は発話に最も近く，話者を指し示す文字列が優先的に選ばれており，entity に限らない．

全ての発話者の性別，別名，そして短い説明文も一覧として提供されているが，本タスクでは利用しない．

4　提案手法

提案手法では，システムの第一段階で文章内の各発話に mention を関連付け，第二段階で mention 情報を利用した発話のクラスタリングを行う．第一段階は Muzny らのふるいを参考に設計したルールベースの手法だが，外部の発話者一覧が利用できない設定で動くよう変更を加えている．第二段階については，クラスタリングに役立つ特徴量の重み付けを行う距離関数を固有名詞を mention に取る発話を手がかりに学習した後，k-means クラスタリングを行う．

4.1　前処理

データセットに対し，まず発話抽出を行い，発話と通常の文章との区別を行った。次に，Stanford CoreNLP [11] を利用して文分割とトークン分割を行い，各トークンへの品詞タグ付けおよび依存構造解析を行った．最後に Stanford CoreNLP で人名を登場人物として認識し，異なる人名として認識される同一人物の重複を認める形で登場人物リストを構築した．

4.2　発話-mention マッピング

本システムの第一段階では，Muzny ら [3] のシステムの第一段階と同様のルールベースのふるいを設けることで，発話と mention を関連付ける．なお，上位のふるいで mention が当たらなかった発話についてのみ，下位のふるいにかけられる．以下に上位のふるいから各ふるいの簡易的な説明を記する．

Trigram 一致　発話-Mention-動詞，などの trigram で，mention が登場人物リストの人物名，あるいは代名詞の場合，発話の mention として採用．

係り受け解析　発話を示す動詞の名目上の主語 (nominal subject) が登場人物リストの人物名，代名詞，あるいは家族を示す名詞の場合，mention として採用．

単一 mention 検出　当該段落の非発話文中に mention が一つのみの場合，mention として採用．

発話パターン検出　以前の発話に特定のパターンに合致する文字列があり，かつ登場人物もしくは家族を示す名詞が含まれる場合，mention として採用．

段落最終 mention 紐づけ　発話が段落の最後にある場合，前文の最後の mention を採用．

会話パターン　発話同士の会話パターンを検出した場合，mention を交互に関連付ける．

弱い会話パターン　会話パターンの制約を弱くし，同様に mention を関連付ける．

なお，最初の 2 つのふるいは既存手法では正解の登場人物リストを用いるが，提案手法では前処理で獲得した登場人物リストを用いることに注意されたい．

4.3 発話クラスタリング

システムの第二段階では，第一段階で得た mention 情報を活用し，以下のステップを通じて発話のクラスタリングを行う．

Step 1: Entity-as-Mention 発話選択　発話全体の集合 Q_A のうち，紐付けられた mention が登場人物リストに含まれている entity である発話 q_n を選び，部分集合 Q_{EM} に加える．

Step 2: 特徴量抽出　各発話 q_n について，発話特徴ベクトル $\boldsymbol{V}_n$ を得る．ここで，$\boldsymbol{V}_n$ は 6 種の特徴ベクトル $\boldsymbol{F}$ の連結であり，

$$\boldsymbol{V}_n = \left[\boldsymbol{F}_n^{\text{emb}}; \boldsymbol{F}_n^{\text{wc}}; \boldsymbol{F}_n^{\text{gen}}; \boldsymbol{F}_n^{\text{near}}; \boldsymbol{F}_n^{\text{in}}; \boldsymbol{F}_n^{\text{men-emb}}\right]$$

となる。特徴ベクトルの設計は以下の通りとなる．

$\boldsymbol{F}_n^{\text{emb}}$: 文埋め込み表現 同一話者の発話は内容が類似すると考えられるため，発話中の単語の単語分散表現の平均を用いる（ベースライン）．

$\boldsymbol{F}_n^{\text{wc}}$: 単語数 長い（短い）発話を行う話者を捉える特徴量． 文章の単語数（文長）として計算．

$\boldsymbol{F}_n^{\text{gen}}$: Gender F-measure 話者の性別を捉えるための特徴量．話者の性別を判断する指針となる F-measure [12] を計算する．$\alpha_{\text{freq.}}$ を名詞の頻度，$\beta_{\text{freq.}}$ を形容詞の頻度，$\gamma_{\text{freq.}}$ を前置詞の頻度，$\delta_{\text{freq.}}$ を冠詞の頻度， $\epsilon_{\text{freq.}}$ を代名詞の頻度， $\zeta_{\text{freq.}}$ を動詞の頻度， $\eta_{\text{freq.}}$ を副詞の頻度， $\theta_{\text{freq.}}$ を間投詞の頻度とする場合，

$$\begin{aligned}\boldsymbol{F}_n^{\text{gen}} = (&\alpha_{\text{freq.}} + \beta_{\text{freq.}} + \gamma_{\text{freq.}} + \delta_{\text{freq.}} + \epsilon_{\text{freq.}} \\ &+\zeta_{\text{freq.}} + \eta_{\text{freq.}} + \theta_{\text{freq.}} + 100) \cdot 0.5\end{aligned} \quad (1)$$

と計算される．

$\boldsymbol{F}_n^{\text{near}}$: 至近 entity の Bag-of-Words 話者や聴者の候補となる人名を捉えた特徴量．具体的には，発話 q_n と同一の段落内，かつ前後 2 文以内の文章に含まれるの Bag-of-Entities とする．

$\boldsymbol{F}_n^{\text{in}}$: 発話内 entity の Bag-of-Words 発話内に含まれる entity は話者自身ではなく，かつ聴者である可能性があると考えられるためこれを用いた．具体的に発話 q_n および mention についての Bag-of-Entities とする．

システム	Precision	Recall	F_1	Accuracy
Muzny et al.[a] [3]	84.6	68.3	75.6	-
Yeung et al.[bc] [9]	-	-	-	52.5
O'Keefe et al.[b] [8]	-	-	-	43.7
提案手法	71.5	56.7	62.5	56.7

a: 正解の登場人物リストが与えられているタスク設定
b: アノテーションの範囲が異なるデータセット
c: 教師あり学習

表 1: Mention 抽出の実験結果．

$\boldsymbol{F}_n^{\text{men-emb}}$: mention 埋め込み表現 代名詞や一般名詞の場合もある mention の意味的な近さを捉えるための特徴量．事前に学習した単語分散表現を $\boldsymbol{X}$，mention を m_n とする場合，mention 中の単語の単語分散表現の平均を用いる．

Step 3: 距離学習　本研究のクラスタリングは，同じ話者の発話を同じクラスタに収めることが目標である．ここで，同一話者の類似性を捉えた特徴ベクトルの次元に重みをかけることで，類似していることが望ましい発話同士の類似性を上げることができる．したがって，次元の重みをかけるための距離行列を学習することで性能の向上が見込める．本手法では，Q_{EM} の発話を訓練データとして，Mahalanobis Metric Learning for Clustering (MMC) [12] の距離行列 $\boldsymbol{L}$ を学習する．

Step 4: クラスタリング　全ての発話に対し，k-means クラスタリングを行う．クラスタ数は前処理の段階で得た登場人物リストに含まれている entity 数とし，発話 q_n に対するデータは $\boldsymbol{V_n L}^T$ とする．

5 実験

実験では，Emma データセットに対して提案手法を適応し，発話の話者同定を行う．本システムは発話と mention を結びつける第一段階と発話のクラスタリングを行う第二段階に分かれており，前者の mention 導出の性能が低いと後者に悪影響を及ぼす．よって，実験 1 では第一段階の mention 抽出性能を検証し，実験 2 でシステム全体の性能を検証する．

なお，システムの第二段階で用いた単語分散表現は，Wikipedia 2014 + Gigaword 5 データセットで学習された GloVe [13] の埋め込み表現（50 次元）[1] を用いた．提

[1] https://nlp.stanford.edu/projects/glove/

$\boldsymbol{V}_n$	MUC Prec.	MUC Rec.	MUC F1	B^3 Prec.	B^3 Rec.	B^3 F1	$CEAF_e$ Prec.	$CEAF_e$ Rec.	$CEAF_e$ F1	Avg. F1
$\boldsymbol{F}_n^{\text{emb}}$	72.54	68.75	70.59	25.34	9.03	13.32	4.11	15.28	6.48	30.13
$\left[\boldsymbol{F}_n^{\text{emb}}\, \boldsymbol{F}_n^{\text{wc}}\right]$	64.57	61.19	62.83	24.89	4.93	8.23	3.41	12.68	5.38	25.48
$\left[\boldsymbol{F}_n^{\text{emb}}; \boldsymbol{F}_n^{\text{wc}}; \boldsymbol{F}_n^{\text{gen}}\right]$	67.33	63.80	65.52	25.66	5.51	9.08	3.76	13.97	5.92	26.84
$\left[\boldsymbol{F}_n^{\text{emb}}; \boldsymbol{F}_n^{\text{wc}}; \boldsymbol{F}_n^{\text{gen}}; \boldsymbol{F}_n^{\text{near}}\right]$	79.14	75.00	77.01	37.80	7.75	12.86	7.52	27.96	11.86	33.91
$\left[\boldsymbol{F}_n^{\text{emb}}; \boldsymbol{F}_n^{\text{wc}}; \boldsymbol{F}_n^{\text{gen}}; \boldsymbol{F}_n^{\text{near}}; \boldsymbol{F}_n^{\text{in}}\right]$	**80.52**	**76.30**	**78.35**	38.89	9.75	15.60	7.55	28.07	11.91	35.29
$\left[\boldsymbol{F}_n^{\text{emb}}; \boldsymbol{F}_n^{\text{wc}}; \boldsymbol{F}_n^{\text{gen}}; \boldsymbol{F}_n^{\text{near}}; \boldsymbol{F}_n^{\text{in}}; \boldsymbol{F}_n^{\text{men-emb}}\right]$	79.44	75.29	77.31	**44.70**	**10.19**	**16.60**	**8.64**	**32.12**	**13.62**	**35.84**

表 2: 教師なし話者同定の実験結果．Avg. F1 は MUC，B^3，$CEAF_e$ の F1 を平均した指標．

案手法の実装にあたり，距離学習の実装及びクラスタリングには，python の metric-learn[2] および scikit-learn[3] のライブラリを使用した．

5.1 実験 1: Mention 抽出性能検証

表 1 に提案システムの第一段階の mention 抽出の性能と既存手法の性能を示す．Muzny ら [3] のシステムは正解の登場人物リストが与えられるタスク設定で解いているため，これを用いない本システムの実装は性能が落ちるが，O'Keefe らのルールベース手法と比べると十分な性能を有している．

5.2 実験 2: システム全体の性能検証

この実験では，提案システムにより生成されたクラスタを評価する．ここで行われたクラスタリングは，同一の話者 entity による発話を同じクラスタに集めているため，共参照解析の評価手法を使うことが適切である．表 2 では，それぞれの発話特徴ベクトルを用いた場合の提案システムを，MUC，B^3，CEAF_e それぞれの尺度で評価したものを示す．

$\boldsymbol{F}_n^{\text{wc}}$ を除き，特徴量が追加されるごとクラスタリングの性能が上がることが確認できる．全ての特徴量を追加した発話特徴ベクトルを用いると，ベースラインの $\boldsymbol{V}_n = \boldsymbol{F}_n^{\text{emb}}$ に比べて Avg. F1 が 5.71 上昇した．

6 おわりに

本研究では，文学作品に含まれる発話の話者同定を教師なしの設定で行うタスクを新たに提案し，距離学習に基づく発話のクラスタリングで話者同定を行う手法を提案した．提案手法は，発話と mention を紐付ける部分タスクで良好な性能を示し，システム全体としても単純な発話内容に基づくベースラインを上回る結果を得た．また，発話の特徴ベクトルについては，有用な特徴量を追加することによる性能向上が認められた．今後は，大規模なデータセットを用いた提案手法の性能評価を行うとともに，本タスクに対して更に有効な手法の研究を進める予定である．

謝辞 本研究は JSPS 科研費 16H02905 の助成を受けたものです．

[2] https://github.com/metric-learn/metric-learn
[3] https://scikit-learn.org/stable/

参考文献

[1] J. Lee and C. Y. Yeung. An annotated corpus of direct speech. In *LREC*, 2016.

[2] J. Li, M. Galley, C. Brockett, G Spithourakis, J Gao, and B Dolan. A persona-based neural conversation model. In *ACL*, pp. 994–1003, 2016.

[3] G. Muzny, M. Fang, A Chang, and D Jurafsky. A two-stage sieve approach for quote attribution. In *EACL*, 2017.

[4] B. Pouliquen, R. Steinberger, and C. Best. Automatic detection of quotations in multilingual news. In *RANLP*, 2007.

[5] K. Glass and S. Bangay. A naive salience-based method for speaker identification in fiction books. In *PRASA*, 2007.

[6] M. S. C. Almeida, M. B. Almeida, and A. F. T. Martins. A joint model for quotation attribution and coreference resolution. In *EACL*, 2014.

[7] D. K. Elson and R. McKeown, K. Automatic attribution of quoted speech in literary narrative. In *AAAI*, 2010.

[8] T. O'Keefe, S Pareti, R. Curran, J, I Koprinska, and M. Honnibal. A sequence labelling approach to quote attribution. In *EMNLP-CoNLL*, 2012.

[9] C. Y Yeung and J Lee. Identifying speakers and listeners of quoted speech in literary works. In *IJCNLP, short*, 2017.

[10] H. He, D Barbosa, and G Kondrak. Identification of speakers in novels. In *ACL*, 2013.

[11] D. Manning, C, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL, System Demonstrations*, 2014.

[12] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPSs*, 2002.

[13] Jeffrey P., Richard S., and Christopher D. M. Glove: Global vectors for word representation. In *EMNLP*, 2014.