BCCWJ を対象としたパターンマッチによる End-to-End 発話者分類

銭本友樹 [1]　古俣慎山 [2]　宇津呂武仁 [1]
[1] 筑波大学大学院 システム情報工学研究群 [2] 筑波大学 理工学群 工学システム学類

概要

小説中の発話文の発話者がどの登場人物かを分類する発話者分類タスクは，小説や登場人物の分析において重要なタスクである．発話者分類は，発話文の抽出，発話者の抽出，同一人物のクラスタリングの 3 つのタスクを行う必要があるが，これらを全て行った日本語小説を対象とした先行研究は存在しない．また評価データが共有されていないため，手法の比較が困難であった．そこで本研究では，第三者が利用可能な『現代日本語書き言葉均衡コーパス』の話者情報アノテーションデータを対象としたパターンマッチによる End-to-End 発話者分類を行い，その有用性と限界について分析する．

1　はじめに

小説中の発話文の発話者がどの登場人物かを分類する発話者分類タスクは，小説や登場人物の分析において重要なタスクである．発話者分類は一般的に，発話文の抽出，発話者の抽出，同一人物のクラスタリングの 3 つのタスクを行う必要がある．

日本語小説を対象とした発話者分類の先行研究には，地の文を利用した研究 [1] と口調を利用した研究 [2, 3] が存在する．Du ら [1] の研究では，青空文庫の小説 4 編を対象として，地の文を利用したパターンマッチによる発話者抽出を試みている．この手法では 0.72 という高い正解率で発話者を抽出できているが，評価に利用した発話文の数は 161 文と少なく，同一人物を指す発話者のクラスタリングは行っていない．石川ら [2] と Zenimoto ら [3] の研究では，ライトノベル中の発話文を対象として，口調の類似性を利用した発話者分類を試みている．しかしながらこれらの手法は，分類先となる発話者の口調が事前情報として必要であり，任意の発話者への分類には対応していない．したがってこれらの日本語小説を対象とした先行研究では，発話文の抽出，発話者の抽出，同一人物のクラスタリングの 3 つのタスク全てを行っておらず，またそれぞれが独自に作成した非公開の評価データを利用しているため，手法の比較が難しいという問題がある．

表 1: 訓練データ中の発話文の両端記号の統計

両端記号	発話文	非発話文
「」	159,493	8,098
()	2,252	3,612
『』	758	3,302
◇	503	1,741
“”	122	1,902
無記号	7,693	—

そこで本研究では，第三者が利用可能な『現代日本語書き言葉均衡コーパス』[1)]（以下 BCCWJ）の大規模な話者情報アノテーションデータ [4] を対象として，発話文の抽出，発話者の抽出，同一人物のクラスタリングの全てを行う End-to-End 発話者分類を試みる．本研究で使用したコードは以下の url[2)] から利用することができる．

2　データセット

BCCWJ 発話者アノテーションデータには，合計 2,845 作品，174,969 文の発話者がアノテーションされた発話文が含まれている．この全体の 2,845 作品のうち，60%を訓練データ，20%を検証データ，20%をテストデータとして扱う．それぞれの小説数と発話文数の統計は付録の表 7 に示す．

3　発話文の抽出

日本語小説中の発話文の特徴を捉えるために，訓練データにおける発話文の両端記号の統計と，同じ記号で囲まれた非発話文の統計を調査した (表 1).

1）https://clrd.ninjal.ac.jp/bccwj/
2）https://github.com/Zeni-Y/speaker-classification

表 2: 発話文抽出の性能比較

両端記号	ルールベース			BERT		
	P	R	F1	P	R	F1
「」	95.5	99.8	97.6	98.9	96.8	97.8
()	49.1	99.3	65.7	93.5	98.8	96.0
『』	16.9	100	28.9	83.4	88.5	85.9
〈〉	12.4	100	22.0	87.9	92.1	89.9
""	2.4	81.8	4.7	36.4	36.4	36.4
無記号	—	—	—	26.5	20.9	23.3
全体	90.4	96.3	93.2	96.7	94.0	95.3

この表の結果から，特定の記号で囲まれた文字列を発話文として抽出する方法では，多くの非発話文を誤って抽出してしまうことがわかった．またこの方法では，両端が特定の記号で囲まれていない発話文を抽出することも不可能である．そこで本研究では，BERT[5] を用いた系列ラベリングによる発話文抽出モデルを構築し，強調や引用表現などの非発話文を誤抽出しない発話文抽出を試みる．

3.1 実験

以下の 2 つの発話文抽出モデルを比較する．

ルールベース 訓練データでの出現頻度の高い 5 種類の両端記号 (「」，()，『』，〈〉，"") によって囲まれた文字列を全て発話文として抽出する．

BERT 事前学習済み言語モデルである東北大版の BERT-base[3] モデルを使用する．訓練データで学習を行い，検証データでの損失が最小のモデルをテストデータでの評価に用いる．

3.2 結果

表 2 に，テストデータを対象とした発話文抽出における Precision，Recall，F1 値を示す．ここでの両端記号全体での評価スコアにはマイクロ平均を用いた．この表から，全ての記号において BERT モデルはルールベースモデルよりも高い Precision と F1 値となる一方で，Recall ではルールベースモデルよりも低い値となることがわかった．また""で囲まれた発話文や，無記号の発話文の抽出性能は BERT モデルでも低いことがわかった．全体の性能としては，「」で囲まれた発話文がほとんどであったために，ルールベースモデルと BERT モデルでの性能に大きな違いはないが，BERT モデルの方がより少ない誤抽出での発話文抽出が可能であることがわかった．

3) https://github.com/cl-tohoku/bert-japanese

4 発話者の抽出

Du ら [1] の手法を参考にし，発話文の周囲の地の文や前後の発話文を利用する 21 種類のパターンを定義し，これらのパターンを利用して発話者の抽出を行う．表 3 に，適用条件ごとのパターン名とその優先順位を示す．これらのパターンが複数適用可能な場合は，表 3 に示す優先順位の最も高いパターンを適用する．パターンマッチを行うための文章の形態素解析や係り受け解析には GiNZA [4] の ja_ginza_electra モデルを利用する．

4.1 人名の抽出

まず以下の 2 つの手法を用いて，地の文から人名の単語及び単語列を抽出する．

人を表す固有名詞の抽出

GiNZA による固有表現抽出によって，「Person」，「Position-Vocation」，「Nationality」，「Name-Other」タグで検出された単語及び単語列を人名として抽出する．

人を表す普通名詞と代名詞の抽出

人を表す単語が網羅された辞書（以下人名単語辞書）を作成し，この辞書中の単語を人名として抽出する．人名単語辞書は，日本語 wordnet[6] から人を表す単語を自動で大規模に収集した後に，人手で不足していた単語の追加と不要な単語の削除を行うことで作成した．

その後，「〜さん」や「〜先生」などの敬称や複合名詞をまとめて人名として扱うために，抽出した単語との係り受け関係が「compound」，「nmod」，「nummod」である連続した単語も含めた全体を，最終的な人名として抽出する．

4.2 発現を表す述語の判定

4.1 節で抽出した人名の係り先である述語が，発言を表す述語かどうかを判定する．発言を表す単語が網羅された辞書（以下発言述語辞書）を作成し，この辞書中の単語ならば発言を表す述語（以下発言述語），それ以外を発言を表さない述語（以下非発言述語）とする．発言述語辞書は，日本語 wordnet[6] から発言を表す単語を自動で大規模に収集した後に，人手で不足していた単語の追加と不要な単語の削除を行うことで作成した．

4) https://megagonlabs.github.io/ginza/

表 3: 適用条件ごとのパターン一覧（パターン名：優先順位）

地の文の位置	主格-発言述語		主格-非発言述語		主辞		その他
	主格	目的格	主格	目的格	主辞	目的格	
同じ	E-S：1	E-S-O：2	I-S：7	I-S-O：8	R-S：13	R-S-O：14	PA1：19
前	E-P：3	E-P-O：4	I-P：9	I-P-O：10	R-P：15	R-P-O：16	PA2：20
後ろ	E-N：5	E-N-O：6	I-N：11	I-N-O：12	R-N：17	R-N-O：18	PM：21

4.3 パターンの適用

4.1 節で抽出した人名の依存関係，4.2 節で判定した述語の種類，それらの人名と述語を含む地の文の位置の 3 つの要素から，表 3 の左の表に示すように 18 種類のパターンを定義する．

例えば図 1 中の発話文「ありがとう」では，同じ行に人名の主格「太郎」を含み，その係り先には発言述語「言った」が存在することからパターン E-S が適用され，「太郎」が発話文「ありがとう」の発話者として抽出される．また発言述語に係る人名の目的格「花子」が存在することから次の発話文「どういたしまして」にはパターン E-S-O が適用され，「花子」が発話文「どういたしまして」の発話者として抽出される．

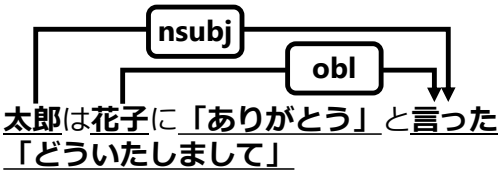

図 1: 人名の主格と目的格を発話者とする例

また図 2 中の発話文「こんにちは」では，前の行で人名の主辞が存在することからパターン R-P が適用され，「佐藤」が発話文「こんにちは」の発話者として抽出される．

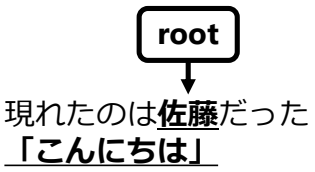

図 2: 人名の主辞を発話者とする例

地の文から発話者の抽出を行えない場合には，以下に示す 3 種類のパターンを適用する．

PA1 or PA2 連続した発話文において，N 番目の発話者が不明で，N±2 番目の発話者が判明しているとき，その N±2 番目の発話者を N 番目の発話者とする．N-2 番目が判明していれば PA1，N+2 番目が判明していれば PA2 を適用する．

PM 他のどのパターンも適用できないとき，小説全体で最も抽出回数の多い人名を発話者とする．

5 同一人物のクラスタリング

4 節の手法で抽出した人名のうち，同一人物を指す人名をクラスタリングする．まず抽出した人名を，「固有名詞」，「接頭辞（例：Mr.，Ms.）」，「接尾辞（例：さん，君）」，「その他（例：先生，将軍）」の 4 つの要素に分解する．そして，「固有名詞」か「その他」のどちらかが一致している人名同士を，同一人物を指す人名としてクラスタリングする．

6 実験

発話者分類はテストデータを対象として行い，クラスタリング評価と人名一致正答率の 2 種類の評価を行う．クラスタリングの評価指標としては，Cuesta-Lazaro らの研究 [7] と同様に B^3 の Precision，Recall，F1 値 [8] を利用する．人名一致正答率では，5 節の手法によって抽出発話者と正解発話者が同一人物と判定された発話文の割合を評価する．またこの際，発話者分類の性能のみを評価するために，分類対象となる発話文はオラクルデータとして与えられているものとする．

7 結果

表 4 に，テストデータ中の 3 つの小説作品単体での発話者分類結果と，小説全体でのマクロ平均と標準偏差を示す．また，テストデータの小説全体の Precision, Recall のヒストグラムを付録の図 3 に，人名一致正答率のヒストグラムを付録の図 4 に示す．表 4 の全体の F1 値から，クラスタリング結果としては，あるクラスタのうち平均して 58.3%の発話文が実際に同じ人物の発話文であり，実際に同じ人物の発話文のうち平均して 49.3%の発話文が同じクラスタに属しているという結果となった．一方で，人名一致正答率は 37.3%と低く，抽出した人名がそのまま正解発話者であることは少ないことがわかった．

続いて小説ごとの分類結果について分析する．

表 4: 発話者分類結果

サンプル ID	発話文数	正解 発話者数	予測 発話者数	Precision	Recall	F1	人名一致 正答率
LBj9_00104	29	2	5	**93.9**	44.6	60.5	51.7
LBf9_00026	80	2	5	54.2	**81.6**	65.1	51.3
LBp9_00033	171	4	17	55.0	39.3	45.8	**8.8**
全体	—	—	—	58.3 ± 14.3	49.3 ± 16.5	52.2 ± 13.4	37.3 ± 18.3

表 5: パターンごとの被覆数（率）

地の文 の位置	主格-発言述語		主格-非発言述語		主辞		その他
	主格	目的格	主格	目的格	主辞	目的格	
同じ	1,328(4.0%)	93(0.3%)	596(1.8%)	57(0.2%)	73(0.2%)	34(0.1%)	PA1:8,387(25.1%)
前	3,740(11.2%)	159(0.5%)	3,760(11.3%)	281(0.8%)	213(0.6%)	449(1.3%)	PA2:3,818(11.4%)
後ろ	5,003(15.0%)	94(0.3%)	4,074(12.2%)	90(0.3%)	299(0.9%)	253(0.8%)	PM:559(1.7%)

表 6: パターンごとの人名一致正答数（率）

地の文 の位置	主格-発言述語		主格-非発言述語		主辞		その他
	主格	目的格	主格	目的格	主辞	目的格	
同じ	860(64.8%)	3(3.2%)	338(56.7%)	4(7.0%)	37(50.7%)	6(17.6%)	PA1:2,365(28.2%)
前	1,392(37.2%)	30(18.9%)	1,476(39.3%)	35(12.5%)	56(26.3%)	53(11.8%)	PA2:1,239(32.5%)
後ろ	2,626(52.5%)	24(25.5%)	1,736(42.6%)	6(6.7%)	96(32.1%)	34(13.4%)	PM:115(20.6%)

表 4 から，サンプル ID が LBj9_00104 の小説では，Precision が高く Recall は低いという結果となった．これは，正解発話者に対して予測発話者の種類が多くなり，同じ人物の発話文が異なるクラスタに分類されてしまったからである．一方でサンプル ID が LBf9_00026 の小説では，Recall が高く Precision は低いという結果となった．これは，予測発話者のほとんどが一人の人物に分類されてしまったからである．また，サンプル ID が LBp9_00033 の小説では，F1 値が高い一方で，人名正答率が著しく低くなった．これは抽出された人名が「僕」や「姉さん」などの代名詞であったために，正解発話者との同一人物判定が失敗してしまうからである．これらの問題の解決には，小説全体での登場人物の特定や，発話文中の口調や二人称の利用，共参照解析による代名詞と固有名詞の同一人物判定の利用が考えられる．

7.1 パターンごとの有用性の分析

パターンごとの有用性について分析する．表 5 にパターンごとの被覆率を示し，表 6 にパターンごとの人名一致正答率を示す．被覆率に着目すると，地の文の位置では，後ろの地の文，前の地の文，同じ行の地の文の順に被覆率が高く，後ろの地の文が最も適用されることが多いことがわかった．一方で PA1 と PA2，PM の被覆率は合計 38.3%となり，多くの発話文が他の発話文の分類結果に依存していることがわかった．人名一致正当率に着目すると，地の文の位置では，同じ行の地の文，後ろの地の文，前の地の文の順に人名一致正答率が高く，同じ行の地の文が最も信頼性が高いことがわかった．地の文の構造に着目すると，主格-発言述語，主格-非発言述語，主辞の順に被覆率と人名一致正答率が高く，発言述語の有用性が高いことが示された．一方で，目的格を使ったパターンは全体的に人名一致正答率が低くあまり有用ではないことがわかった．

8 終わりに

本研究では，BCCWJ 中の小説を対象とした End-to-End 発話者分類に取り組んだ．発話文抽出においては，特定の記号で囲まれた文字列を抽出するルールベースモデルが不十分であることを示し，多様な形式の発話文を高い精度で抽出可能な BERT を利用した系列ラベリングモデルを提案した．発話者分類では，周囲の地の文を利用したパターンマッチによる発話者の抽出により，同じ発話者の発話文の半数程度を適切に分類できることを明らかにした．またパターンごとの被覆率と人名一致正答率を詳細に分析し，パターンごとの有用性を明らかにした．

謝辞

本研究は科研費 21H00901 の助成を受けたものである．また本研究では，国立国語研究所のプロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー・小磯花絵）および日本学術振興会・科学研究費補助金「会話文への発話者情報の付与によるコーパスの拡張」（15H03212）の成果データを利用した．

参考文献

[1] Du Yulong, 白井清昭. 小説からの自由対話コーパスの自動構築. 言語処理学会第 25 回年次大会発表論文集, pp. 623–626, 2019.

[2] 石川和樹, 宮田玲, 小川浩平, 佐藤理史. 口調ベクトルを用いた小説発話の話者推定. 研究報告自然言語処理（NL）, pp. 1–8, 2019.

[3] Yuki Zenimoto and Takehito Utsuro. Speaker identification of quotes in Japanese novels based on gender classification model by BERT. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation. Association for Computational Lingustics, 2022.

[4] 山崎誠, 宮嵜由美, 柏野和佳子. 小説会話文への話者情報付与. Technical report, 国立国語研究所, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, 2019.

[6] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet. In Proceedings of the 7th Workshop on Asian Language Resources (ALR7), pp. 1–8, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[7] Carolina Cuesta-Lazaro, Animesh Prasad, and Trevor Wood. What does the sea say to the shore? a BERT based DST style approach for speaker to dialogue attribution in novels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5820–5829, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Amigó Enrique, Gonzalo Julio, Artiles Javier, and Verdejo and Felisa. A comparison of extrinsic clustering evaluation metrics based on formal constraints. In Inf. Retr., pp. 12(4):461–486, 2009.

表 7: BCCWJ 中のデータ統計

—	小説数	発話文数
訓練データ	1,707	104,711
検証データ	569	35,417
テストデータ	569	34,841
合計	2,845	174,969

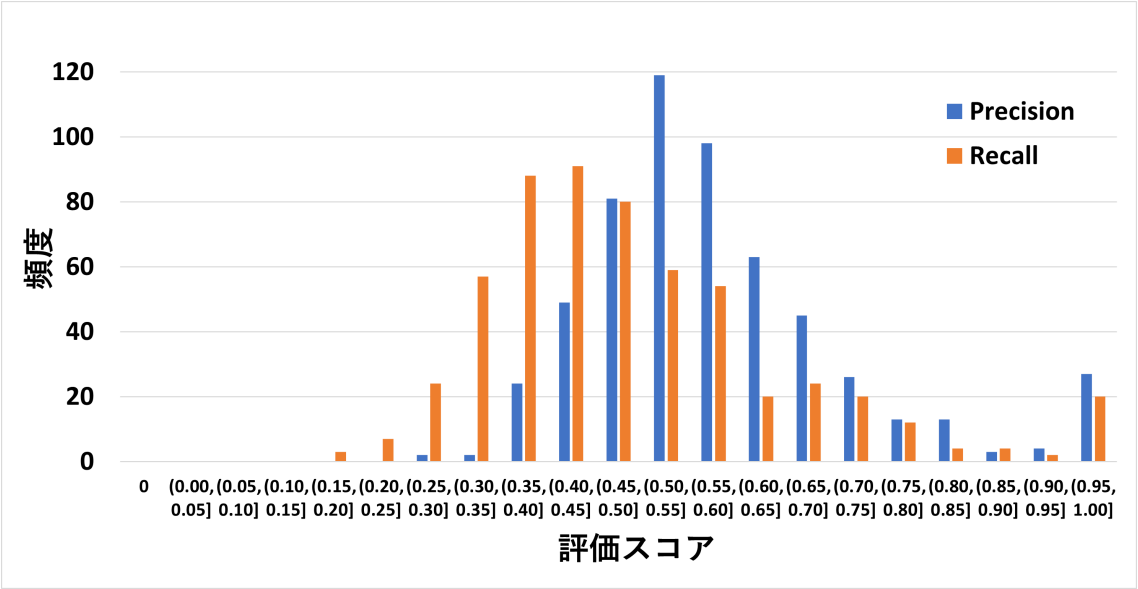

図 3: テストデータ全体の Precision, Recall のヒストグラム

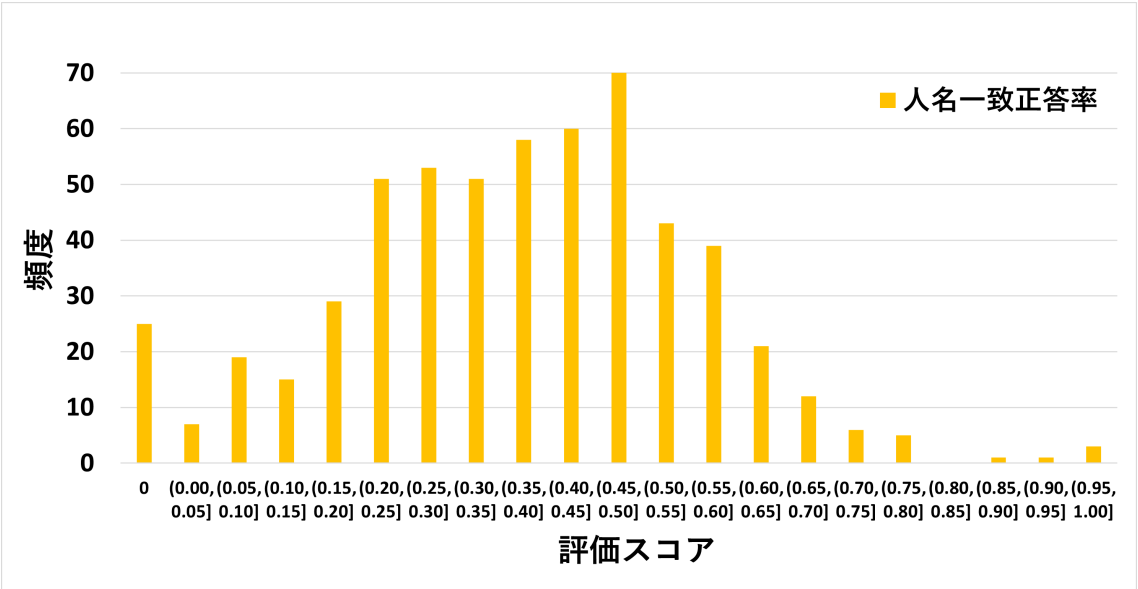

図 4: テストデータ全体の人名一致正答率のヒストグラム