小説からの自由対話コーパスの自動構築

Du Yulong　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1710137,kshirai}@jaist.ac.jp

1 はじめに

ユーザと雑談する自由対話システムは近年注目を集めている研究トピックである．自由対話システムの構築や評価には対話コーパスが欠かせない．しかし，複数人による対話を収録し，それを書き起こす作業を必要とするため，対話コーパスの構築は作成コストが高く，大規模なコーパスを整備するのが難しいという問題がある．

この問題に対し，Twitter のようなマイクロブログから複数人のやり取りを擬似対話とみなして自動収集する試みが行われている [6]．本研究では，小説から対話を抽出することで対話コーパスを自動構築することを試みる．小説には登場人物の台詞があり，複数の人物による連続した台詞は対話とみなせる．小説における対話は作例であり，完全に自然な対話とは言えないが，作家は自然に発生する会話を想定しているため，自然な対話に近い性質を有すると考えられる．一方，大量の小説を対象に対話を収集すれば，大規模な対話コーパスを低コストで構築できるという利点がある．

対話コーパスの構築を目的とする場合，小説から連続した台詞を抽出するだけではなく，それぞれの台詞を発っした話者を特定し，話者の情報を付与した発話 (台詞) の列を抽出する方が望ましい．He らは，英語の小説を対象に，台詞の発話者を推定する手法を提案した [2]．小説の中から明示的に書かれている台詞の話者をパターンマッチで抽出し，その小説における全ての台詞の話者の候補とする．次に，それぞれの台詞に対し，話者の候補の中から適切な話者を決定するランキングモデルを学習している．ただし，正解の話者がタグ付けされた訓練データを必要とする．一方，小林は，物語をシーンごとに分割する手法を提案した [3]．既存の辞書などを利用して場所，時間，人物候補を抽出し，これらの 3 種類の候補の異なり数を基準としてシーンを分割する．この研究では，小説をシーンに分割することで，人物が特定の場面に存在するか否かの「入退場情報」を決めることができるが，台詞とその発話者の対応は決めていない．

本論文では，これらの先行研究を踏まえつつ，小説における台詞の中でも特に連続して出現する台詞の話者を特定する手法を提案し，話者の情報を付与した自由対話コーパスを自動構築する方法を提案する [1]．He らの手法 [2] と異なり，正解の話者がタグ付けされた正解データを必要としない手法を提案する．

2 提案手法

提案手法の処理の流れは以下の通りである．まず，与えられた小説のテキストに対し，本文以外のメタデータの削除や文分割などの前処理を行った後，台詞と登場人物を抽出する．次に，抽出した個々の台詞に対し，その話者を特定する．最後に，連続している台詞を抽出し，話者の情報とともに対話コーパスを出力する．

2.1 台詞の抽出

台詞の抽出はパターンマッチにより行う．具体的には，以下の 6 種類の開括弧と閉括弧の組で囲まれた文字列を台詞として抽出する．.+は任意の文字列にマッチすることを表す．

「.+」　『.+』　（.+）　《.+》　——.+」　——.+』

2.2 登場人物の抽出

西原と白井の手法 [4] にならい，以下の 2 つの方法で登場人物を抽出する．

- 固有表現抽出による手法
  固有表現抽出によって「人名」として検出された単語もしくは単語列を登場人物として抽出する．具体的には，CaboCha[1] によって「Person」とタグ付けされた単語列を抽出した．
- シソーラスによる手法
  シソーラスで人物に相当する意味クラスを持つ名詞を登場人物として抽出する．本研究では，シソーラスとして日本語語彙大系を利用し，「人名」「人」のカテゴリに含まれる語を抽出した．

2.3 台詞の話者の特定

小説の全ての台詞の話者を特定する．その手続きを図 1 に示す．なお，本研究では台詞を以下の 2 つに分

[1] http://taku910.github.io/cabocha/

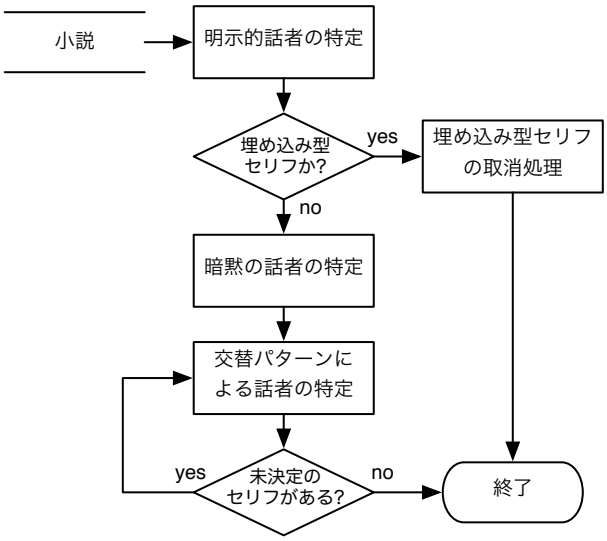

図 1: 台詞の話者の特定

類する．

埋め込み型台詞 文の途中に出現する台詞．

(例) 父は「ありがとう」と言った．

独立型台詞 それだけで一文を構成する台詞．

2.3.1 明示的な話者の特定

まず，台詞を発言した人物が明確に記述されている話者を「明示的な話者」と定義し，これを特定する．明示的な話者は，表 1 に示す $\mathrm{PE}_1 \sim \mathrm{PE}_7$ の「明示的話者特定パターン」を用いて特定する．これらのパターンにおいて，U は 2.1 項で検出された台詞を，P は 2.2 項で検出された登場人物を表す．J はトピックを表す係助詞である[2]．

一方，SV は発言を表す動詞 (speech verb) である．インターネット上のシソーラスや類語辞典を参考に，SV に該当する 1,171 個の動詞のリストを作成した．図 2 にその一部を示す．

ささやく，しゃべりたてる，しゃべる，叫ぶ， 仰しゃる，言う，語る，告白する，説明する

図 2: 発言を表す動詞 (抜粋)

明示的話者特定パターンでは，指定するパターンにマッチしたとき，台詞 U の話者を人物 P と特定する．パターン PE_3，PE_4，PE_7 では，台詞 U_1 と U_2 の話者をそれぞれ P_1，P_2 と特定する．パターン $\mathrm{PE}_1 \sim \mathrm{PE}_7$ をこの順序で適用し，最初にマッチしたパターンによって話者を特定する．また，同じパターンで複数の登場人物にマッチしたときは，台詞 U との距離が一番近い登場人物を特定する．表 1 の第 2 列は，それぞれのパターンにマッチする台詞と話者の例である．

2.3.2 埋め込み型台詞の取消処理

明示的話者特定パターンによって話者を特定できない台詞が埋め込み型台詞のとき，それを台詞として検出した処理を取り消し，台詞ではないものとみなす．予備調査の結果，表 1 のパターンで話者を検出できない埋め込み型台詞は，括弧で囲まれていても台詞ではない場合がほとんどであったためである．一方，独立型台詞については後続の処理で話者を特定する．

2.3.3 暗黙の話者の特定

台詞を発言したことが明確に記述されてはいないが，暗黙的に示されている話者を「暗黙的な話者」と定義し，これを特定する．暗黙的な話者は，表 2 に示す $\mathrm{PI}_1 \sim \mathrm{PI}_4$ の「暗黙的話者特定パターン」を用いて特定する．U，P，J は台詞，登場人物，トピックを表す係助詞である．これらのパターンは，基本的に，台詞の前後の文に出現する登場人物を話者として特定している．ただし,「は」「ては」などトピックを表す係助詞の前に出現する人物を優先して特定する．すなわち，パターン $\mathrm{PI}_1 \sim \mathrm{PI}_4$ をこの順序で適用し，最初にマッチしたパターンによって話者を特定する．同じパターンで複数の登場人物にマッチしたときは台詞 U との距離が一番近い登場人物を特定する．

2.3.4 話者交替パターンによる話者の特定

明示的話者特定パターンと暗黙的話者特定パターンでも話者を特定できない場合は,「話者交替パターン」を用いる．複数の台詞が連続して出現するとき，その台詞の話者は交替することが多い．表 3 に示す話者交替パターンはこの性質を利用して話者を特定するものである．パターン PA_1 において，(台詞) は小説中の台詞を，(話者) は既に特定された台詞の話者を表す．いま，U_2 の話者は「人物 A」と決まっているが，その次の台詞 U_3 の話者は決まっていない．このとき，U_2 の直前に出現する台詞の話者が「人物 B」と特定されていれば，U_2 から話者が交替すると仮定し，U_3 の話者を「人物 B」と特定する．なお，台詞が連続している場合だけでなく，間に短い文が挿入されているときでも，同様に話者交替のパターンが適用できると考えられる．そのため，U_2 と U_3 は連続した台詞だが，U_1 と U_2 の間には台詞以外の文が存在してもよいものとする．PA_1 における U_1 と U_2 の間の $*$ は任意の文の出現を許すことを表す．つまり，U_1 は U_2 の前に出現する一番近い台詞である．

[2]「は」「も」「では」「には」「や」「が」のいずれか．

表 1: 明示的話者特定パターン

PE_1: P J $*$ U と $*$ SV	父$_P$ は$_J$ 小さく「ありがとう」$_U$ と 言った$_{SV}$ 。
PE_2: U と 、 P J $*$ SV	「ありがとう」$_U$ と 、 父$_P$ は$_J$ 小さく 言った$_{SV}$ 。
PE_3: P_1 J $*$ P_2 に $*$ U_1 と $*$ SV U_2	太郎$_{P_1}$ は$_J$ 花子$_{P_2}$ に 笑顔 で「ありがとう」$_{U_1}$ と 言った$_{SV}$ 。 「どういたしまして」$_{U_2}$
PE_4: U と 、 P_1 J $*$ P_2 に $*$ SV U_2	「ありがとう」$_{U_1}$ と 、 太郎$_{P_1}$ は$_J$ 花子$_{P_2}$ に 笑顔 で 言った$_{SV}$ 。 「どういたしまして」$_{U_2}$
PE_5: P J $*$ SV 。 U	彼女$_P$ は$_J$ 心から 謝罪$_{SV}$ し た 。 「ごめんなさい」$_U$
PE_6: U P J $*$ SV 。	「ごめんなさい」$_U$ 彼女$_P$ は$_J$ 心から 謝罪$_{SV}$ し た 。
PE_7: U_1 P_1 J $*$ P_2 に $*$ SV U_2	「ありがとう」$_{U_1}$ 太郎$_{P_1}$ は$_J$ 駅 で 花子$_{P_2}$ に そっと 伝えた$_{SV}$ 。 「どういたしまして」$_{U_2}$

P: 人物，J: トピックを表す係助詞，U: 台詞，SV: 発言を表す動詞.

表 2: 暗黙的話者特定パターン

PI_1: $*$ P J $*$ U	刑事$_P$ は$_J$ 走り出し た 。 「そいつ を 捕まえろ！」$_U$
PI_2: U $*$ P J $*$	「財布 を 落とし ま し た よ」$_U$ 青年$_P$ が$_J$ 立っ て い た 。
PI_3: $*$ P $*$ U	気づい た の は 刑事$_P$ だっ た 。 「そいつ を 捕まえろ！」$_U$
PI_4: U $*$ P $*$	「財布 を 落とし ま し た よ」$_U$ 声 の 主 は 青年$_P$ だっ た 。

P: 人物，J: トピックを表す係助詞，U: 台詞.

パターン PA_2 も PA_1 と同じ考えにしたがって話者を特定する．また，PA_1 と PA_2 はこの順序で適用する．

話者交替パターンは，全ての台詞の話者が特定されるまで，あるいは話者交替パターンによって話者を特定できる台詞がなくなるまで，繰り返し適用する．

2.4 対話の抽出

小説中の全ての話者を特定した後，連続する台詞を対話として抽出する．小説内の会話では台詞の間に地の文が出現することもあるので，台詞間に出現する文の数が 2 以下のときは連続した台詞であるとみなす．対話を抽出する際には，その話者の情報も一緒に抽出する．さらに，今後の予定として，話者を A，B といった記号 (話者 ID) に置き換えることを検討している．図 3 は実際に抽出された対話の例である．各行は (今後付与する予定の) 話者 ID，小説から特定された話者，台詞を示している．なお，† のついた話者は解析誤りで，正しくは「子供」である．

表 3: 話者交替パターン

PA_1

(話者)	(台詞)
人物 B	U_1
人物 A	$*U_2$
?	U_3

? → 人物 **B**

PA_2

(話者)	(台詞)
?	U_1
人物 A	U_2
人物 B	$*U_3$

? → 人物 **B**

3 評価実験

3.1 話者特定手法の評価

2.3 項で述べた台詞の話者を特定する手法を評価する．青空文庫 [5] に掲載されている小説の中から 4 編を選び，台詞とその話者を人手でタグ付けし，これをテストデータとする．評価基準は適用率と正解率とする．適用率は，小説に含まれる台詞のうち，(正解，不正解を問わず) 話者を特定できた台詞の割合と定義する．正解率は，話者を特定できた台詞のうち，正しく話者を特定できた台詞の割合と定義する．テストデータに対する実験結果を表 4 に示す．

適用率は全ての小説で 1 となった．つまり，全ての台詞について話者を特定できた．一方，正解率は，3 つ

A	男	『なぜ黙ってるの。』
B	彼女	『まだ本当にわづかしか経ちませんのね、結婚してから。』
A	男	『さうだなァ。』
A	男	『たった三年にしかならないんだな。けれども、俺たちはいろ／＼苦労したなァ。』
B	彼女	『本当にね。』
A	男	『さうだ、一日々々いろ／＼なことに疲らされなやまされ苦しまされても、二年はもう過ぎたんだからな。』
B	彼女	『坊や。』
B	彼女†	『うま／＼／＼／＼。』
A	男	『もう少しの間だ。』
B	彼女	『さうね、私たちは働きませうね。』
A	男	『さ、あとを片づけよう。そして寝よう、明日は早く起きようぢゃないか。』

図 3: 抽出された対話の例

表 4: 話者特定手法の評価結果

タイトル	台詞数	適用率	正解率
晩餐	16	1.00	0.88
端午節	57	1.00	0.82
象牙の牌	61	1.00	0.74
さいかち淵	27	1.00	0.37
(全て)	161	1.00	0.72

の小説については高い値となった.『さいかち淵』の正解率が低いのは，登場人物がニックネームで表現されていて，固有表現抽出やシソーラスを用いた手法では人名と認識できなかったためである．登場人物リストのような情報があれば，登場人物抽出の再現率が上がり，話者を正確に特定できるようになると考えられる．

解析誤りの主な原因について述べる.『晩餐』において，1 つの文に 2 つの台詞が存在する場合があり，このときに話者を特定することができなかった．これに対しては，明示的話者特定パターンを追加することで解決できる可能性がある.『端午節』と『象牙の牌』では，複数の台詞が連続するときに，同じ話者が 2 回連続で台詞を発言するときがあり，このときに誤った話者が特定された．話者交替パターンでは，話者が台詞毎に必ず交替することを仮定していたが，例外的にそうでない場合があるので，対処が必要である．また，現在の話者交替パターンは対話の参加者が 2 名であることを仮定しているため，3 名以上の人物が会話している場面では正しい話者を認識できない．

3.2 対話コーパスの構築

青空文庫 [5] に掲載されている小説から対話を抽出した．結果を表 5 に示す．話者特定の適用率は 0.917 であった．表中の「対話数」は，連続して出現しかつ全ての台詞の話者を特定できた対話の数である．平均して 10 個程度の発話からなる 19,000 件の対話を抽出し，比較的大規模な対話コーパスを自動構築できた．

表 5: 対話コーパス構築の実験結果

小説数	4,838
うち台詞を検出できた小説数	3,267
うち対話を抽出できた小説数	2,209
対話数	19,492
対話における発話 (台詞) の総数	201,391
一対話当たりの平均発話数	10.3

4 おわりに

本論文では，対話コーパスを自動構築するために，小説における台詞の話者を特定し，連続する台詞を話者の情報とともに対話として抽出する手法を提案した．

今後の課題として，対話コーパスの整備のために，小説中の登場人物名を話者 ID へ置き換える処理を実装することが挙げられる．この際，代名詞が抽出されたとき，その先行詞を特定し，同一人物に対して同じ話者 ID を与える工夫が必要である．また，表 4 の実験は，提案手法の検討に用いた小説をテストデータとして用いるため，クローズドテストとなっている．新しいテストデータを用いたオープンテストによる評価が必要である．

参考文献

[1] Yulong Du. 小説からの対話コーパスの自動構築. 修士論文, 北陸先端科学技術大学院大学, 3 2019.

[2] Hua He, Denilson Barbosa, and Grzegorz Kondrak. Identification of speakers in novels. In *Proceedings of ACL*, pp. 1312–1320, 2013.

[3] 小林聡. 場・時・人に着目した物語のシーン分割手法. 情報処理学会研究報告, Vol. 2007-NL-179, No. 47, pp. 25–30, 2007.

[4] 西原弘真, 白井清昭. 物語テキストを対象とした登場人物の関係抽出. 言語処理学会第 21 回年次大会, pp. 626–631, 2015.

[5] 野口英司（編）. インターネット図書館青空文庫. はる書房, 2005.

[6] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *HLT-NAACL*, pp. 172–180, 2010.