物語テキストにおけるキャラクタ関係図自動構築

神代 大輔† 高村 大也†† 奥村 学††
†東京工業大学 大学院総合理工学研究科 ††東京工業大学 精密工学研究所
kamishiro@lr.pi.titech.ac.jp {takamura,oku}@pi.titech.ac.jp

1 はじめに

近年，著作権の切れた物語を電子化し無料公開する青空文庫[1]や，電子化された小説を WEB 上で販売する e-novels[2]が登場し，またネット上でアマチュアが小説を公開するサイトが林立するなど，電子化された物語テキストは着実に我々の身近なものになってきている．また電子ブック端末の登場や，各種モバイル端末での小説閲覧システムの普及など，近い将来，物語テキストの舞台が紙から電子媒体に移っていくことが予想される．このような現状において，電子データの特性を活かせば，紙媒体では不可能だった，読者にとって有用な情報をテキストから自動提示することが可能となってくる．そこで本研究では，物語テキストから登場キャラクタ同士の関係を推定，関係図を自動構築することを目的とする．物語を読む際に，読んでいるページまでのキャラクタの関係図を参照できれば，長い物語を読む際の理解の手助けとなる．また未読の物語の関係図を参照し，読むか否かの意思決定に役立てるなどの使い方も期待できる．本研究においては，発話しているエンティティをキャラクタ候補，キャラクタ候補の中で他候補との関係の繋がりが密な候補をキャラクタと定義する．まず物語中に含まれる各発話文について話し手と聞き手を同定し，その会話の中身に暗黙的に示された情報から，キャラクタ候補同士の関係を推定する．1 発話文について話し手ノードと聞き手ノードの間に 1 つのエッジを張ることにより，グラフを形成する．ノードの次数から，関係から孤立したもの，極端に薄く繋がっているものを除外できるようにし，最終的なキャラクタの関係図を出力する．

2 関連研究

英語の童話から話し手を特定している研究として，Zhang ら [1] のものがある．これは固有名詞抽出技術と品詞列パターンから話し手となりうるエンティティ候補を特定し，発話と同じパラグラフ内の候補から，発話の直前，なければ直後のものを話し手と選定している．本研究では日本語の物語テキストを対象とする．日本語の物語テキストでは段落の区分けが明確でなく，また第 3 節で記述する独立型発話において，話し手が発話の前後どちらに明示されるかが曖昧である．本研究ではこの話し手の明示位置を考慮することによって，精度の高い話し手同定を行う．

日本語の物語テキストを対象とした関連研究として，馬場ら [2] のものがある．馬場らは，物語テキストから人物の属性情報を抽出する手法，人物同士の関連度を計算して相関図を描画する手法を提案した．本研究ではキャラクタ単独の属性ではなく，会話に着目することによって，キャラクタ間の関係を推定する．

[1]http://www.aozora.gr.jp
[2]http://www.so-net.ne.jp/e-novels/top.htm

発話から人物間の関係推定を行った関連研究としては，西原ら [3] のものがある．西原らは，音声会話，Web 上のチャット，掲示板への書き込み，メールのやりとりなどを対象に，人物間の仲の良さと上下関係を推定する手法を提案した．本研究において考える人物関係は物語中のものであり，こうした現実での会話からみられる特徴とは異なっていると考えられる．本研究では実際に物語中の会話の中身を捉えることにより，現実ではなく物語内のキャラクタ間の関係を推定する．

3 話し手・聞き手同定

本節では物語中の各発話文について，話し手となるエンティティと聞き手となるエンティティを同定する．

3.1 発話の型分類

本研究では物語中の発話を，その形態によって大きく以下の二つの型に分類する．

組込み型発話 地の文の中に発話が含まれた形態のもの．文脈に因らず，ある程度発話者の明示される場所が拘束されている．『**A**は「○○○」と言った.』『「□□□」と**彼**は答えた.』のように，含まれた地の文の中に発話者が明示されていることが多く，含まれていない場合も『**A**が走ってきた．息を切らせながら「○○○」と言った.』のように，発話の前方に発話者が明示されていることが多い．『「□□□」と聞こえた．言ったのは**A**だった.』のように発話後方に明示されることは稀である．

独立型発話 発話が一文を構成するもの．『「○○○」**A**は言った.』や『**A**は言った.「□□□」Bはそれを聞いて頷いた.』のように，発話の前後方どちらに発話者が明示されるかが周辺文脈に因り，曖昧である．

話し手同定においては，この曖昧性を解消することが必要となる．

3.2 提案手法

前述のように，話し手同定においては，主に独立型の発話について当該発話の正しい話し手を特定することが必要となる．機械学習によってこの問題を解決する際に問題となるのが，正解となる話し手の種類・数が物語ごとに異なるため，直接正解となるエンティティをラベルとして学習することができない点である．

そこで本研究では，図 1 のように，発話から相対的にどの位置に話し手が明示されているかを示したラベルを当てることによって，間接的に話し手を同定する手法を提案する．“前方 1”，“後方 1”，“前方 2”，“後方 2”，“前方 3”，“後方 3”の 6 つのラベルを用意し，教師あり学習の手法により，発話周辺の文脈によって話し手がどの位置に明示されるかを学習する．発話周辺の話し手正解のうち，発話に最も近い位置にある話し手正解を指すラベルを正例として与え，残りのラベ

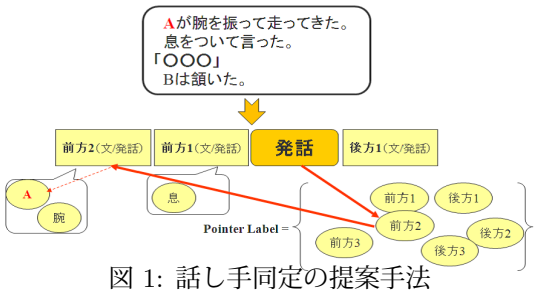

図 1: 話し手同定の提案手法

表 1: 使用した素性

構造	発話か地の文か．発話の型 (独立型，前後方組込型，後方組込型，前方組込型)
地の文内のエンティティの助詞・係り先	以下の条件を満たすエンティティが含まれるか否か.『助詞:は/が/も　係り先:発言系動詞』『助詞:は/が/も　係り先:意思系動詞』『助詞:は/が/も　係り先:動詞 or サ変名詞』『助詞:は/が/も』『係り先が発話を跨ぐ』『条件なし』
自律度	各自律度を持つエンティティが含まれるか否か.

ルを負例として与える．これは発話から近い位置にあるエンティティの方が，遠い位置にあるものに比べて，たまたま地の文に現れるわけでなく，その発話の主体を明示するために記述されることが多いためである．使用する素性を表 1 に示す．それぞれについて，対象発話とその前方後方 3 文についての素性を入れる．出力ラベルが指した先が地の文の場合，文内から，あらかじめ規定した助詞と係り先の優先度によってエンティティを選択する．出力ラベルが指した先が発話の場合，その発話と同じ話し手を当該発話に与える．ただし指した先の発話が当該発話を指し返している場合，分類スコアが次点のラベルを選択するものとする．

聞き手同定は，話し手同定の結果を踏まえて行う．物語中で一度以上発話をした話し手候補集合とその発話の中から，当該発話に最も近くかつ当該発話の話し手と違うものを聞き手として選択する．前後 3 文以内に条件を満たすものがない場合は聞き手不在の発話とする．また，発話中に話し手候補に対しての呼びかけが行われている場合，対象エンティティを聞き手として選択する．表記されたエンティティが呼びかけかどうかは，句読点または閉じ鍵括弧の直前に表記されているかどうかで判別する．なお条件を満たす場合でも呼びかけとなっていない場合 (並列明示など) があるが，数が少ないのでここでは考慮に入れないものとする．

3.3　自律度の導入

発話と話し手の関係について,『彼が頷いた.「○○○」顔が笑っている.』のような例を考える．この例において，“彼”と“顔”のどちらが話し手かに着目した場合，助詞・係り先の情報のみから考えると，“エンティティ *1* が ⇒ 頷く”と“エンティティ *2* が ⇒ 笑う”の二つを比べることになり，判別は困難となってしまう．話し手同定の精度を高めるためには，“顔”のような喋らなさそうなエンティティと，“彼”のような喋りそうなエンティティを判別する指標を導入する必要がある．しかし物語テキストにおいては動物や物体が喋ることもあるので，単純に概念辞書とマッチングをするだけでは解決できない．本研究では，喋りそうなエンティティは文章全体で動詞に多く係っていると仮定し，地の文内のエンティティに関して，

自律性 物語全体での動詞/サ変名詞との係り比率を，0.2 刻みで，0.0〜1.0 までの 5 段階

出現数 物語全体での出現数を，1〜2，3〜4，5〜6，7〜8，9 以上，の 5 段階

に分け，自律度をこの 2 つのペア 25 種類で表し，素性に加えることによってこれを解決する．

3.4　話し手・聞き手同定実験

青空文庫から収集した物語テキスト 23 作品で実験を行った．総発話数は 1118 で，内訳は組込み型発話:551，独立型発話:567 だった．

3.4.1　ベースライン

発話の型ごとに，以下の順番で文を探索する．

- 組込み型：発話が含まれる文内 ⇒ 前方 1 文 ⇒ 前方 2 文 ⇒ 後方 1 文 ⇒ 後方 2 文
- 独立型：後方 1 文 ⇒ 前方 1 文 ⇒ 後方 2 文 ⇒ 前方 2 文 ⇒ 後方 3 文 ⇒ 前方 3 文

エンティティを発見したら，あらかじめ規定した助詞と係り先の優先度により，エンティティを選択する．なお組込み型発話において，エンティティの係り先が発話を跨いでいる場合，優先度を上げる．これは『A がそう言うと B は「○○○」と答えた.』のように文内に複数のエンティティが存在する場合，係り先が当該発話を跨ぐものの方が話し手になりやすいと考えられるためである．探索範囲内に条件を満たすエンティティが存在しない場合，最も近い発話を選び，その発話と同じ話し手を当該発話に与える．独立型発話が 4 つ以上連続して出現した場合，2 番目の発話，3 番目の発話に，それぞれ 4 番目の発話，1 番目の発話と同じ話し手を与える．

3.4.2　提案手法を用いた実験

ラベルつきデータ 23 作品すべてについて，各テスト作品以外の 22 作品を訓練に用い，正解率を算出した．実験結果を表 2，3 に示す．データ内の各発話について，それぞれ正しい話し手，聞き手を同定できた数を計測し，正解率を算出した．

話し手同定においては，組込み型についてはベースラインが上回っているが，独立型については提案手法の方が良い正解率を出すことができた．これは組込み型については前述のように話し手の明示場所がかなり拘束されているので単純に文内を探索した方が良いが，独立型についてはそれが曖昧であるため学習による効果が現れてくるためと考えられる．以降，聞き手同定，関係推定実験で使う話し手同定システムにおいては，組込み型についてはルールベース手法を，独立型については提案手法を用いた．

聞き手同定においては，話し手として正解を与えた場合は約 8 割の聞き手を当てることができたが，システム出力を与えた場合は 5 割強に留まり，話し手同定の精度に強く依存した．呼びかけを用いた場合は，用いない場合と比べて，話し手として正解を与えた場合は 0.4 %，システム出力を与えた場合は 2.0 %正解率が上がっており，呼びかけが有効であることを確認した．

なお，話し手同定，聞き手同定をかけたとき，話し手・聞き手の両方を正解したものの割合は，49.9 %だった．

表 2: 話し手同定 正解率 (%)

	組込み型	独立型	全発話
ベースライン	87.86	50.00	70.97
提案手法	83.69	57.73	72.61

表 3: 聞き手同定 正解率 (%)

	呼びかけなし	呼びかけあり
話し手正解を与えた	80.8	81.2
システム出力を与えた	54.7	56.7

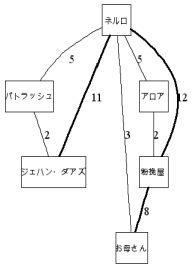

図 2: 相関図出力 (次数上位 6 まで)

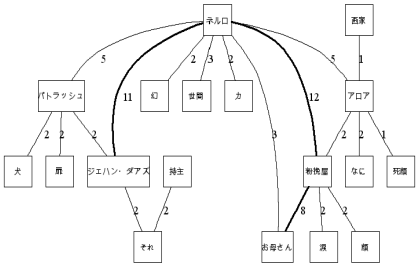

図 3: 相関図出力 (すべて)

3.5 相関図出力

話し手・聞き手同定の結果をもとに，キャラクタの相関図を出力する．物語作品に話し手・聞き手同定をかけ，キャラクタ間の 1 発話を 1 本のエッジとしてグラフを形成した．なお相関図においてはエッジの方向性は重要でないため，図の見やすさを考慮し無向グラフとした．次数が閾値以上のものを出力した．“フランダースの犬”についての出力結果例を図 2，3 に示す．主人公であるネルロが一番次数が大きく，他もおおむね主要なキャラクタは次数が大きくなっていることが確認できる．

4 関係推定

前節の手法で同定した物語中のキャラクタ同士の会話から，友好敵意関係，目上目下関係の二軸について，キャラクタ間の関係を推定する．

4.1 提案手法

キャラクタ間の関係に人手でラベルを付けて特徴を学習する教師あり学習の手法をとった場合に問題となるのが，ラベル付けのコストである．1 つの物語についてのキャラクタ間のリンク数は多くないため，大量の物語データを読んでラベルを付けなければならない．また物語内での発話の量が少ない短い物語からは，学習できる特徴が少なくなるため，必然的にある程度以上の文章量の物語にラベルを付けることが求められる．両条件を考えると，非常にコストが大きくなってしまい，ラベル付けは実用的でない．本研究ではこの問題を，ラベルつきデータを用いず，大量のラベルなしデータを用いることで解決する．

日本語において，人間同士の関係を表す強い指標となるものに人称表現がある．たとえば話し手が一人称として“わたくしめ”を使っていれば話し手が相手よりも目下の関係にあることがわかり，二人称として“貴様”を使っていれば話し手が相手に対して敵対的な態度をとっていることがわかる．現実の会話においてこうしたあからさまな人称が現れることは稀であるが，物語内ではよく見られる．そこで本研究では，こうした

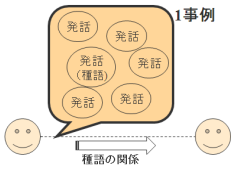

図 4: 事例の作り方 (キャラクタリンク単位)

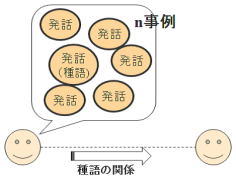

図 5: 事例の作り方 (発話単位)

一人称二人称による限られた特徴のみを用いてキャラクタ間の関係ラベル付けを自動で行い，そこから各々の関係についての特徴を学習する手法を提案する．

Wikipedia[3]内の「一人称」,「二人称」の項目に記された人称から，関係を表すと考えられるものを厳選し，種語とする．ラベルなしデータに対し話し手・聞き手同定を行い，発話のやりとりのある 2 キャラクタ間にリンクを想定し，*Character1* から *Character2* への発話集合 SP の中に種語が条件を満たす形で含まれていれば，*Character1* から *Character2* へのキャラクタリンクにその関係のラベルを与える．*Character1* から *Character2* に貼られたラベルが“目上⇒目下”の場合は *Character2* から *Character1* に“目下⇒目上”のラベルを，*Character1* から *Character2* に貼られたラベルが“目下⇒目上”の場合は *Character2* から *Character1* に“目上⇒目下”のラベルを追加で付与する．ただし矛盾する関係の種語が同時に含まれた場合は，ラベルを付与しない．

このデータから“友好”ラベルを正例，“敵意”ラベルを負例として学習した友好敵意分類器と，“目上⇒目下”を正例，“目下⇒目上”を負例として学習した目上目下分類器を作成する．友好敵意分類と目上目下分類それぞれについて，キャラクタリンク単位で分類するものと発話単位で分類するものの 2 種類の分類器を作る．

キャラクタリンク単位 図 4 のように，発話集合 SP 内の発話をまとめて 1 つの事例として扱う．学習の際には固有名詞，種語を含む n-gram は除外する．分類の際には Support Vector Machine(SVM) を用い，*Character1* から *Character2* への発話集合 SP をまとめて 1 事例として分類器を適用する．確信度として SVM の出力する分離超平面からの距離を用い，確信度が閾値よりも小さい事例については“どちらでもない”クラスに分類する．

発話単位 図 5 のように，SP に含まれる個々の発話について，キャラクタリンクの関係と同じ関係を持つ 1 つの事例とみなして扱う．学習の際には固有名詞，種語を含む n-gram は除外する．分類の際には SVM を用い，*Character1* から *Character2* への発話集合 SP 内の各発話について個別に分類器を適用する．出力される分離超平面からの距離を，正例寄りならば正の値，負例寄りならば負の値として合算し，確信度とする．確信度が閾値よりも小さい事例については“どちらでもない”クラスに分類する．

4.2 関係推定実験

4.2.1 人手ラベル付与データの作成

まず提案手法の評価用に，人手で関係ラベルを付与したデータを用意した．物語テキスト 23 作品において，発話のやりとりがある登場キャラクタ間に人手で関係

[3] http://www.wikipedia.org/

表 4: 人手ラベル付与データの関係ラベル数

友好	敵意	どちらでもない
41	15	226

目上⇒目下	目下⇒目上	どちらでもない
59	59	163

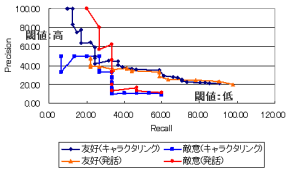

図 6: RP 曲線 (友好敵意分類)

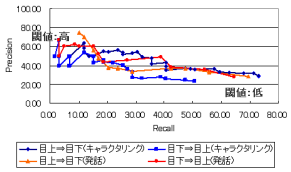

図 7: RP 曲線 (目上目下分類)

のラベルを付与した．ラベルは友好敵意分類では“友好”，“敵意”，“どちらでもない”の 3 種類，目上目下分類では“目上⇒目下”，“目下⇒目上”，“どちらでもない”の 3 種類とした．表 4 の数のラベルが貼られた．以下，このデータを人手ラベル付与データとする．提案手法の評価と，提案手法と人手ラベル付与データを用いた教師あり学習手法との比較評価に用いる．

4.2.2 提案手法を用いた実験・考察

ラベルなしデータ 1666 作品に対し，提案手法によって関係のラベルづけと学習を行った．素性として友好敵意分類では unigram+bigram，目上目下分類では unigram+bigram+trigram を用いた．これは予備実験を行ったもののうち一番結果が良かったものである．

次にテストデータとして人手ラベル付与データを用い，提案手法の評価実験を行った．確信度の閾値を変化させたときの精度と再現率を算出した．結果を図 6, 7 に示す．友好敵意分類において，“友好”クラスの方が結果が良い傾向にあった．これは“友好”については比較的手掛かりとなる口調が出やすいが，“敵意”の手掛かりとなるような乱暴な口調は出現する数が少ないことが一因として考えられる．また友好は直接的に示されるが，敵意は直接的に示されにくいことも要因と考えられる．たとえば“シンデレラ”における“母⇒シンデレラ”や，“さるかに合戦”における“さる⇒かに”には“敵意”のラベルが貼られているが，これらは相手に直接的に乱暴な言葉遣いをして敵意が示されているわけでなく，皮肉や，言動ではなく裏で意地の悪い行動をするなどの記述で敵意が表されている．目上目下関係についても，実際の関係と喋るときの振る舞いの違いが影響を及ぼした．身分や属性により関係として目上目下関係にあっても，目下から目上にくだけた調子で喋りかけていたり，目上から目下に敬意をもって話している場合があり，分類を誤ることが多かった．また“目下⇒目上”については，キャラクタリンク単位のものより発話単位のものの方が全体として結果が良かった．これは目上目下分類においては機能語に特徴が偏りやすいので，素性の頻度を考慮していないためにキャラクタリンク単位の場合は平滑化されてしまう特徴が，発話単位の場合は捉えられるためではないかと考えられる．

4.2.3 教師あり学習手法との比較実験

提案手法の有効性を確認するため，提案手法で学習した分類器と，人手ラベル付与データから教師あり学習手法で学習した分類器との，比較実験を行った．ここで，教師あり学習手法について，人手ラベル付与データに話し手・聞き手同定をかけ“友好”，“敵意”，“どちらでもない”の三値分類器，“目上⇒目下”，“目下⇒目上”，“どちらでもない”の三値分類器を学習して交差検定を行うと，“どちらでもない”の出力が殆どを占め，それ以外のラベルについて有効な精度を得られなかった．これは学習する事例数の偏りのためと考えられる．そこで教師あり学習手法についても，“友好”を正例，“敵意”を負例として学習した友好敵意分類器と，“目上⇒目下”を正例，“目下⇒目上”を負例として学習した目上目下分類器を作成し，“どちらでもない”ラベルは直接学習しないこととした．学習の際，発話集合 SP 内の発話をまとめて 1 つの事例として扱い，固有名詞を含む n-gram は除外した．素性は提案手法に合わせ，予備実験を行ったもののうちそれぞれ一番結果が良かったものを用いた．分類の際には SVM を用い，*Character1* から *Character2* への発話集合 SP をまとめて 1 事例として分類器を適用した．確信度として SVM の出力する分離超平面からの距離を用い，確信度が閾値よりも小さい事例については“どちらでもない”クラスに分類した．提案手法，教師あり学習手法それぞれについて，閾値を変化させたときの“友好”，“敵意”，“目上⇒目下”，“目下⇒目上”の精度，再現率の変化を図 8，9 に示す．各関係ラベルにおいて，提案手法の方が教師あり学習手法に比べて高い結果を得ており，提案手法の有効性を確認できる．

提案手法と教師あり学習手法の比較結果

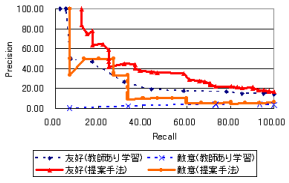

図 8: 友好敵意分類

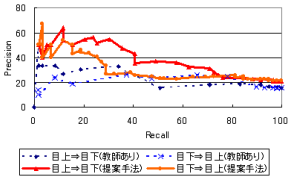

図 9: 目上目下分類

5 まとめと今後の課題

本研究では，物語テキストから登場キャラクタ同士の関係を推定し，関係図の自動構築を行う手法を提案した．その際，物語中に含まれる各発話文についての話し手と聞き手を同定し，その会話の中身から話し手同士の関係を推定した．話し手同定においては，話し手の明示場所を示したラベルを当てることによって，機械学習を用い，精度を上げた．関係推定においては，大量のラベルなしデータから限られた特徴のみを用いてキャラクタ間の関係ラベル付けを自動で行うことで，ラベル付けに関するコストを抑えた．今後の課題としては，話し手同定においてラベル同定後のエンティティ選定の厳選，関係推定において地の文からの情報活用などが挙げられる．

参考文献

[1] Jason Y Zhang, Alan W Black, Richard Sproat. Identifying speakers in children's stories for speech synthesis. EUROSPEECH-2003, pp.2041–2044, 2003.
[2] 馬場 こづえ, 藤井 敦. 小説テキストを対象とした人物情報の抽出と体系化. 言語処理学会第 13 回年次大会発表論文集, pp.574–577, 2007.
[3] 西原 陽子, 砂山 渡, 谷内田 正彦. 発話テキストからの人間の仲の良さと上下関係の推定. 電子情報通信学会論文誌, Vol.J91-D, No.1, pp.78–88, 2008.