映画あらすじ文からの登場人物関係の抽出

服部 純次† 杉本 徹‡
† 芝浦工業大学大学院工学研究科電気電子情報工学専攻
‡ 芝浦工業大学工学部情報工学科

1. まえがき

映画のあらすじを理解するには，その映画の登場人物がどのような人物で，どのような動作を行うのか，といったことを把握する必要がある．中でもストーリーを把握する上で，登場人物どうしの関係がどうなっているか，ということを理解することは極めて重要である．本研究では，映画のあらすじについて書かれた文章を解析し，登場人物どうしの関係を抽出する手法を提案する．この関係は，登場人物が映画中で行う動作から求める．

このような情報の利用方法としてストーリーに基づく映画の推薦や検索が考えられる．人によって好きな映画のストーリーや登場人物の関係や構成はある程度似通っていると考えられる．例えば，三角関係や味方だった者が裏切る，また同僚と恋をするなど，様々なものが考えられる．世の中に存在する映画のうち一人の人が知っている映画はごく一部であり，まだ知らない映画の中から好みに合ったストーリーを持つ映画を推薦することができれば，有益であろう．また現在映画を検索する際には映画名により検索することがほとんどであるが，映画の人物関係に関する情報を与えることで検索することができるようになれば，役に立つ場面も多いのではないかと考えられる．

本研究では，その第一段階として与えられたあらすじ文から登場人物どうしの関係を抽出する手法を提案する．

2. 使用する題材

本研究では「goo 映画」(http://movie.goo.ne.jp/)という Web サイトに掲載されている映画のあらすじ文を題材として使用する．このサイトには，1920～2005 年の間に公開されたおよそ 30000 作品のあらすじ情報が載っており，1 作品あたり平均 700 文字程度，多いものでは 1000 文字を超える長さのあらすじ文が掲載されている．

3. 登場人物情報の抽出

上記の Web サイトから得た映画のあらすじ文を解析し，登場人物の動作を元にして，他の登場人物との関係を示す要約を生成する．要約は，登場人物ごとの動作に着目し，他の人物と共起する文を見つけ，重要だと思われる箇所を自動的に抽出する．ある人物に対して，他の人物との共起の数だけ関係を導けるため，関係を一つに絞らずに抽出でき，複雑な人間関係も表現できる．

Extraction of the Character Relationship from an Outline Text of a Movie
Junji Hattori† Toru Sugimoto‡
† Graduate school of Engineering, Shibaura Institute of Technology
‡ Department of Information Science and Engineering, Shibaura Institute of Technology

3.1 形態素・係り受け解析

取得してきたあらすじ文を MeCab および CaboCha を用いて解析し，文に含まれる単語の品詞や文節間の係り受け関係を認識する．これによって品詞に基づいて人物や動詞を判別したり，動詞の主語を求めたりできるようになる．また MeCab によって人名判断をする際，初期状態では人名に限りがあり，特に外国人の名前などは判断できる数が少ないため，抽出できない名前も多数存在する．そこで，MeCab が利用している IPA 辞書に人名の追加を行うことによって精度が向上する．本研究では暫定的に，実験に使用する作品に登場する人名のみ追加している．

3.2 人物の動作の抽出 [1]

まず登場人物の動作に関する情報，つまり主語と動詞とそれ以外の補足情報（「を」格など）を抽出する．抽出する動作は，「を」格などに人物を含むものに限る．

あらすじの中には 2 種類の動詞の出現パターンが考えられる．1 つ目は「私は車を買った」のような動詞が文の末尾に来る場合，2 つ目は「車を買った山田は…」のような連体修飾句として現れる場合である．

動作を抽出するために，まずあらすじ文中の動詞に着目し，その動詞の主語を求める．そのために動詞に係る文節で助詞が「が」や「は」，「も」などの係助詞であるものを探し，その文節が人名とみなされる名詞を含む場合に主語と定める．もし見つからなかった場合は，その動詞が別の動詞に係っているならば，その主語が今着目している動詞の主語にもなっていると考える．連体修飾句の場合は，動詞が係っている名詞に人名が含まれていれば，その人物を動詞の主語とみなす．

さらに，補足的な情報として同じ動詞に係る文節で「で」，「を」，「に」などの格助詞を持つ文節を取り出す．また補足的な情報を補う情報（「～の」など）も取得する．最後に，求めた主語と動詞，補足情報を組にして抽出する．

4. 登場人物どうしの関係の抽出方法の検討

4.1 人物関係を表す動作表現へのタグ付け

一般に人物どうしの関係は人物の動作表現（手掛かり語と呼ぶ）から判断することができる．どのような動作表現がどのような人物関係を表すか調べるために，goo 映画より取得してきたあらすじ情報の中で，比較的抽出数が多かった 16 作品のあらすじ文に対して，文中に現れる手掛かり語とそれが表す人間関係の種類のタグ付けを行った．関係の種類は，敵仇，恋愛，味方，主従，仕事，同僚，同一，親類，接有の 9 個とした．なお接有とは，「訪ねる」や「会う」といったような，あらすじ文の中で頻出する言葉ではあるが，特定の関係を導くには至らないような言葉で共起している人物どうしの関係を表す．

付けたタグの総数は 128 個であり，一作品内での重複を除くと 125 個であった．また手掛かり語は全部で 114 種類あった．それらを人物関係の種類ごとに分類したものを表 1 に示す．なお関係の () 内の数字は付けたタグの数を示す．

表 1

関係	手掛かり語
敵仇 (59)	「倒す」，「殺す」，「討つ」など 54 種類
味方 (26)	「救う」，「頼む」，「諫める」など 25 種類
接有 (14)	「訪ねる」，「会う」，「訪れる」など 11 種類
恋愛 (13)	「恋する」，「思う」，「惹く」など 13 種類
主従(9)	「命じる」，「従える」など 6 種類
同一(4)	「改名する」，「改める」，「となる」
同僚(1)	「組む」
仕事(1)	「依頼する」
親類(1)	「娘と名乗る」

4.2 抽出方法 1 ：完全一致による関係の抽出

表 1 のような人物関係の種類ごとに分類された手掛かり語のリストを利用することにより，与えられたあらすじ文から登場人物どうしの関係を抽出することができる．2 人の人物が主語または「を」格，「に」格などに共起する文において手掛かり語が使われていた場合，その語が示す人物関係を抽出する．

この抽出方法の再現率を見積もるために，今回タグ付けをした 16 作品の中から 1 作品ずつ順に取り出し，取り出した作品を除く 15 作品に含まれる手掛かり語を用いて，その作品に現れる関係を抽出できるか調べる．その結果，全部で 125 個の関係のうち再現されたのは 18 個であり，再現率は 14.4%となった．再現された言葉は，「倒す」や「殺す」，「命じる」などであった．

4.3 抽出方法 2 ：動詞の類似性を考慮した抽出

再現率を向上させるためには，前節で述べた方法を拡張していく必要がある．実際にタグ付けした手掛かり語を見てみると，同じ関係を表す語の中には，「訪れる」と「訪ねる」や「斬る」と「斬り捨てる」など，意味的に似ている語が多く存在していることが分かる．そこで，与えられた文中に手掛かり語と完全一致する語が現れる場合だけでなく，類似した語が現れる場合にも関係抽出を行うことを考える．

たとえば実際にタグ付けの結果，「敵仇」と分類した手掛かり語である「殴る」に対して，その他すべての手掛かり語との類似度を EDR の概念階層上の類似度に基づいて計算してみたところ，類似度の高いものから順に「打つ」，「叩く」，「斬る」，「痛めつける」となった．この 4 つの語はどれも「敵仇」を表す手掛かり語である．このことは，類似度計算による関係の抽出が有効な方法だと示している．しかし，一意に関係が決められない場合もある．たとえば「疑う」を他のすべての手掛かり語と類似度計算すると，上位の 4 語は順に「思う」，「入れる」，「得る」，「知る」となった．これらの関係はどれも「疑う」が分類されている関係「敵仇」とは違っている．また「裏切る」の場合も，類似度計算を行うと，上位から順に「売る」「訪ねる」「訪れる」「追う」「させる」となる．しかし「裏切る」の関係は「敵仇」と分類しているが，上記の 5 つのうち「追う」以外は「敵仇」に分類されていない．

このように，ひとえに類似度だけでは分類が難しいのが現状である．類似度の閾値をどこにするかなど，決定する必要がある．

4.4 問題点とその改善方法

また「拳銃を見舞う」や「助けを求める」など動詞だけでなく「を」格などの句を伴うことによって関係を分類できるものも多い．たとえば「切腹を命じる」では，現段階において「命じる」から関係を「主従」と分類しているが，さらに「切腹を命じる」として比較することで，「敵仇」の関係をともに導ければ，関係の広がりをもたらすことができると考えられる．「逆鱗に触れる」の場合も，「触れる」だけでは様々な関係を推測できるが，「逆鱗に触れる」として考えることでまた別の人間関係を導けると考えられる．

現段階では，このような 2 語以上からなる手掛かり語に関して，類似度の算出を行うことができない．たとえば，「逆鱗」と「触れる」の類似度を別々に計算した後，それらの組み合わせで「逆鱗に触れる」の類似度を計算できそうだが，それではそれぞれの平均を算出するに過ぎず，文の意味を的確にとらえられない場合もありうる．今後そのような算出方法の検討が必要である．

5. まとめと今後の課題

映画のあらすじ文を解析し，登場人物どうしの関係を抽出する手法の提案をした．映画作品によりあらすじ文にそれぞれ特徴があり，抽出できる動作数が一定でないのが現状である．できるだけ多くの情報を抽出するために，動作の抽出の精度向上が必要である．

関係の抽出においても，1 語だけの比較では，タグ付けされた言葉に関しては抽出できるが，タグ付けしていない言葉では，明らかにその精度が低下する．その解決方法として，4.4 での例のように「逆鱗に触れる」など 2 語以上の言葉から抽出することが望ましい．そうした類似度計算の方法を今後検討していきたい．

また，関係の手掛かり語をタグ付けした作品の数がまだ少ないため，タグ付けを行う作品を増やし，手掛かり語を増やしていくことで，精度の向上に努めたい．

さらに，抽出した人物関係の情報を応用した映画推薦や検索の方法についても今後検討していきたい．

参考文献

[1] 服部純次, 杉本徹: "映画あらすじ文からの登場人物情報の抽出", 第 7 回情報科学技術フォーラム, 2008.

[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 (NLP2007) , pp.574-577, 2007.