E-008

映画あらすじ文からの登場人物情報の抽出

Extraction of the Character Information from an Outline Text of a Movie

服部 純次† 杉本 徹‡
Junji Hattori Toru Sugimoto

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，また登場人物どうしの関係がどうなっているか，ということを把握する必要がある．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，登場人物に関する情報を抽出する手法を提案する．ここで登場人物に関する情報としては，登場人物が映画中で行う動作とその人物の属性，および人物間の関係を扱う．

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構成はある程度似通っていると考えられる．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，映画のストーリー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の登場人物情報の抽出を試み，その精度についての評価を行う．

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．要約の生成

上記の Web サイトから得た映画のあらすじ文を解析し，登場人物の動作や属性を中心とした要約を生成する．要約は，登場人物ごとにその動作や属性，その人物の重要度や他の登場人物との関わりなどをまとめた人物フレームの集合として表される．映画「風林火山」のあらすじ文を解析した結果得られる「勘助」という登場人物の人物フレームの一部を表１に示す．

表１．生成された人物フレームの例（抜粋）

勘助	登場回数	18
動作	晴信の家臣になったのだった	
	武田の家老板垣に恩を売り，	
	和議を整えた	
属性	新参者	
共起関係	晴信	5
	武田	2
	由布	1
	謙信	1

このような要約を求める手順を以下に示す．

① 形態素・係り受け解析
② 人物の動作の抽出
③ 人物の属性の抽出
④ 主役の同定
⑤ 人物間の共起の抽出

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．

3.2 人物の動作の抽出

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（目的語など）を抽出する．

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人物フレームに登録する．

† 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

3.3 人物の属性の抽出

次に登場人物の属性情報の抽出を行う．ここで属性とは「美貌の」，「美しい」など人物を修飾している言葉のことであり，職業や性別，年齢などの情報を含む．goo 映画に掲載されているあらすじ文を調べたところ，人物の属性を以下の５つの形式で表していることが多いことが分かった．

① 家老板垣（人名の前に一般名詞がつく）
② 弁護士・杉浦（中黒）
③ ギタリストのノブ（「の」がついた名詞句）
④ 妻であるテス（「である」がつく句）
⑤ クールなナナ（形容動詞や形容詞がつく）

これらの形式で人物を修飾している属性語を抽出して，動作に関する情報と同様に，人物フレームに登録する．

3.4 主役の同定と人物間の共起の抽出

あらすじ文から映画の主役を同定するために，あらすじ文中にそれぞれの人名が動詞の主語として現れる回数を用いて，最も回数が多い人を主役とみなす．

さらに，登場人物間の共起の回数を求める．ここで，ある人物の動作を表す文の中にほかの人物が現れるとき，それらの人物は共起していると言う．人物の共起は，誰と誰が多く関わりを持っているかを表す．

4. 評価

4.1 人物の動作の抽出

「風林火山」，「OUT」，「NANA」，「13 階段」の４作品のあらすじ文を対象として本手法を用いて要約生成したものと，人手で要約生成したものとを比較した．

動作の抽出に関する比較の結果を表２に示す．ここで「動作数」は人手で抽出された（正解の）動作数を表す．

表２．抽出した動作の適合率と再現率

映画名	動作数	適合率	再現率
風林火山	55	88.2%	81.8%
OUT	22	87.0%	90.9%
NANA	16	71.4%	62.5%
13 階段	23	69.6%	69.6%

「風林火山」と「OUT」に関しては，適合率，再現率とも 80～90%という比較的よい結果が得られた．一方，「NANA」と「13 階段」では，適合率，再現率とも 60～70%という低い結果となった．その原因としては，抽出対象の文がかなり長文であったことが挙げられる．長文のため，係り受け解析の結果が間違っていたり，あるいは係り受け解析が正しくできていたとしても，正しく係り先が辿りきれなかったと考えられる．特に主語が文の真ん中あたりに存在する場合に，動作の抽出が正しく行えないことがあった．

4.2 人物の属性の抽出

動作の抽出と同様に属性の抽出に関して比較を行った結果を表３に示す．

表３．抽出した属性の適合率と再現率

映画名	属性数	適合率	再現率
風林火山	21	38.1%	72.7%
OUT	12	41.7%	100.0%
NANA	17	58.8%	62.5%
13 階段	21	66.7%	100.0%

再現率に関しては「OUT」と「13 階段」で 100%であった．「NANA」で再現率が低い値となった原因の一つとして，名前を認識できずに属性を抽出できないことがあった．

一方，適合率は全体的に低い数値となった．属性抽出のための判定条件を見直して，誤抽出を少なくし，適合率を高めることが今後の課題である．

4.3 主役の同定と人物間の共起の抽出

主役の同定に関しては，「風林火山」で「勘助」の主語としての出現回数が 18 回と最も多くなり，主役と同定された．これを含めて４作品すべてで主役である登場人物を正しく同定することができた．

また人物間の共起に関しては，表１に示した「風林火山」の「勘助」において顕著な傾向が見られた．実際，映画の中で「勘助」と「晴信」は強い関わりがあり，共起数の大きさはそれを反映している．しかしあらすじ文がそれほど長くない他の３作品では，人物間の共起数に大きな差が見られなかった．今後，共起数だけでなく，共起の仕方の違いも考慮した抽出方法を考えていきたい．

5. まとめと今後の課題

映画のあらすじ文を解析し，登場人物の動作や属性，共起数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，すべてのあらすじ文に対して，同様に予想した結果が得られていないのが現状である．今後より多くの作品に対して高い精度で情報抽出ができるように手法の改良を進めていきたい．

今後の課題として，登場人物間の関係の種類，たとえば恋人や家族，ライバルなどの関係性を導くことが挙げられる．その方法として，関係の種類を示す言語表現をあらすじ文から収集し，辞書として利用する方法を考えている．

さらに，あらすじの要約を利用した映画推薦，検索を実現するために，要約を照合して類似度を算出する方法についても今後検討していきたい．

参考文献

[1] 照井康太, 舘寛典, 袖山智, 杉本徹, 榎津秀次: "意味の類似性に基づく映画推薦システムの提案", 第 69 回情報処理学会全国大会, 2007.

[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.