小説における場面変化の情報抽出

吉原　亮　　前田　修作　　五十嵐　俊介　　韓　東力
日本大学文理学部　情報システム解析学科

１　はじめに

自然言語処理における小説に関する研究はいくつか行われている。小説を読者に理解しやすいように整理された情報を提供するためである。小説には登場人物や場面など様々な情報がある。

しかし、既存の研究では主に小説の登場人物、すなわち「人」に関する研究がほとんどである。人物相関図や登場人物の特徴を抽出するものなどがあった[1][2][3]。そこで我々は「人」ではなく、今まであまり注目されていなかった、「場面」に注目し、場面の変化をもとに小説の情報を整理していく。

２　本研究の考え方

本研究では場面変化の情報抽出においてパターンマッチング方式を採用している。パターンを決定するのに参考にした文章は芥川龍之介の作品で「魚河岸」「仙人」「蜘蛛の糸」「十円札」の４作品で、青空文庫[1]より取得している。文章の選考は以下の基準をもとに行われている。

・著作権が切れているもの
・場面変化があるもの
・長すぎないもの
・新字新仮名になっているもの

次に、本研究で取り扱われる場面変化の定義を述べる。

・場面変化
　場所や状況が切り替わること
　一枚の挿絵で場景を表わす際の単位
・場所属性
　場所の様子（人・物など）を表わす要素
・場面属性
　場所同士のつながり（位置関係や包含関係など）や場面自体の様子を表わす要素

場所属性と場面属性は場面変化を抽出する際に必要なものである。

次に、場面変化の具体例を一つ上げる。

例：　A さんは学校で授業を終え、駅に向かった。A さんは寄り道することなく自宅に帰った。

図１　場面変化の例

図１は A さんのいる場所が学校から駅になり、駅から自宅になる様子を表したものである。

本研究では会話文や回想部分を考慮していない。その理由は、会話文や回想部分の中だけで話が進んでいってしまい、本人は元の位置から動いていないことがあるからである。

３　システムの流れ

本研究で構築したシステムの全体図は以下のようになっている。図中の完全一致単語辞書と正規表現単語辞書は日本語語彙大系[4]をもとにあらかじめ作成したものである。完全一致単語辞書は単語自身が完全に一致するもので、正規表現単語辞書は単語を正規表現で一致させるものである。

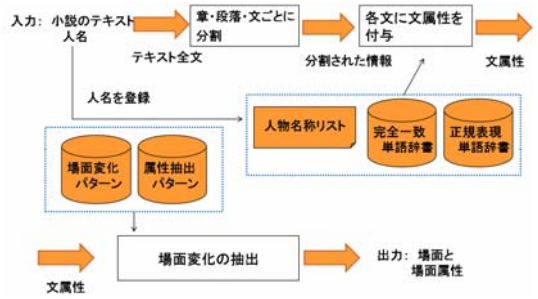


図２　システムの全体図

３．１　入力データの前処理

まずはテキスト形式の小説を開き、人物名を登録する。これにより、人物名リストを作成していく。次に、一文ずつ解析をしていくため、章、段落、文ごとに分ける。それと同時に、会話文を考慮しないので会話文の削除も行う。

３．２　文属性の付与

文章を係り受け解析器CaboCha[2]にかけて、完全一致単語辞書や正規表現単語辞書などを用いて、各文に文属性を付与していく。

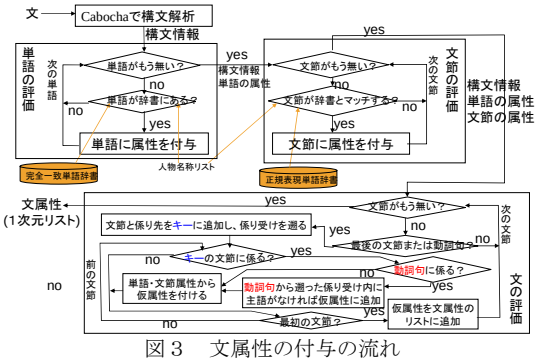


図３　文属性の付与の流れ

[1] http://www.aozora.gr.jp/

[2] http://chasen.org/~taku/software/cabocha/

３．３　場面変化の抽出

付与された文属性をもとに場面変化パターンと属性抽出パターンを適用し、場面変化と属性を抽出する。パターンについては４章で詳しく説明する。

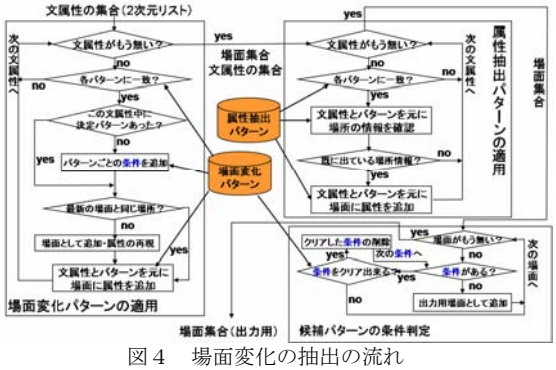


図４　場面変化の抽出の流れ

３．４　システムの出力

場面と場面属性が出力され、どの人物がどの場面にいるのか、どの場面からどの場面へ移動したのかが分かるようになっている。また、一文ずつの詳しい解析結果も見ることができる。

４　場面変化の抽出パターン

本研究では、場面変化箇所を特定するために場面変化パターン、場面中の人物や物を抽出するために属性抽出パターンをあらかじめ作成した。場面変化パターンは「決定パターン」と「候補パターン」にわけられている。すべてのパターンはCaboCha による係り受け関係をもとに作成している。

４．１　場面変化パターン

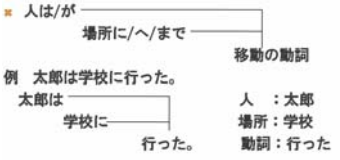
場面変化パターンとは、入力の一文が場面変化の箇所となっているかを判断するためのパターンである。場面変化と場面情報の抽出が、それぞれ前後の文に関係するか否かで、計４種類に分けられる。

・決定パターン・完全
・決定パターン・条件
・候補パターン・完全
・候補パターン・条件

以下にそれぞれ一例ずつを示す。

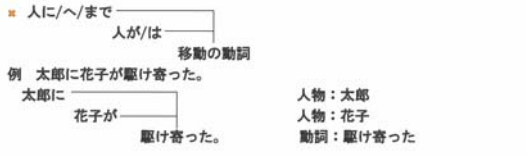
①　決定パターン・完全

場面変化、場面属性ともに抽出に条件がないパターン。



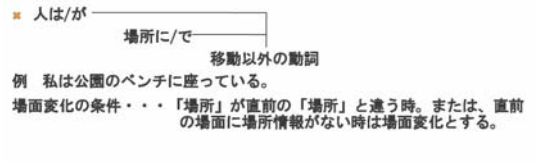
②　決定パターン・条件

場面変化の抽出に条件はないが、場面属性の抽出には条件があるパターン。



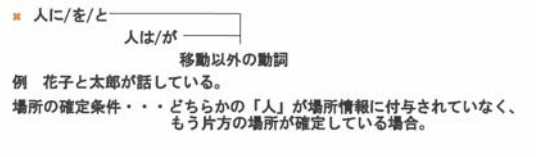
③　候補パターン・完全

場面変化の抽出に条件はあるが、場面属性の抽出には条件がないパターン。



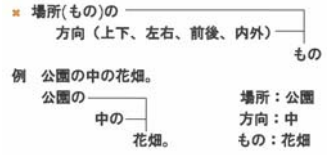
④　候補パターン・条件

場面変化、場面属性ともに抽出に条件があるパターン。



４．２　属性抽出パターン

属性抽出パターンとは、入力の一文が既出場面の情報を持っているかを判断するためのパターンである。場面変化パターンをベースに作られたものが多く、また、場面変化パターンのような区分けはされていない。



上記以外にもいくつかパターンはあるが、ここでは省略する。

５　評価実験

手法の有効性を検証するため、評価実験を行った。この章では実験の方法と結果を詳しく述べる。

５．１　実験方法

実験では、人手で小説の場面変化を確認し、場面変化箇所や、場面変化の場所の抽出を確認した。今回の実験ではクローズドテストとオープンテストの二種類の実験を行った。各実験は第一段階と第二段階に分かれており、第一段階では場面変化箇所の特定をする。場面変化箇所とは、場面の名前を抽出することができないが、そこで場面が変化したことを表わすものである。第二段階では場面変化の場所の名前を抽出する。場面変化の場所の名前とは、小説内で登場する場所の名前のことで、例えば、「学校」や「家」などを表わす。

テストの結果は出力された答えをそのまま人手

で作成した解答と比較し、適合率・再現率・F 値を求める。

人手で作成した解答の集合を N とし、システムの出力された集合を M とする。

再現率（P）＝|M∩N|/ |M| ×１００(%)
適合率（R）＝|M∩N|/ |N| ×１００(%)
F 値＝２PR /（P+R）×１００(%)

本研究では会話部分や回想部分を考慮していないので、その部分は N、M からも削除して計算してある。

平均値は、すべての小説の N の和集合を平均値の N とし、同様にすべての M の和集合を平均値の M とし、上記の計算式で値を求める。

５．２　クローズドテストと結果

パターン選考に用いた４つの作品（「魚河岸」「仙人」「蜘蛛の糸」「十円札」）で精度調査を行った。

表１　第一段階クローズドテスト結果

	再現率	適合率	F 値
魚河岸	６６．７%	４０．０%	５０．０%
仙人	１００%	３７．５%	５４．５%
蜘蛛の糸	６６．７%	３３．３%	４４．４%
十円札	５５．０%	６８．８%	６１．１%
平均値	６２．１%	５１．４%	５６．３%

表２　第二段階クローズドテスト結果

	再現率	適合率	F 値
魚河岸	６６．７%	４０．０%	５０．０%
仙人	６６．７%	２５．０%	３６．４%
蜘蛛の糸	１００%	３３．３%	５０．０%
十円札	５５．０%	６８．８%	６１．１%
平均値	６０．７%	４８．６%	５４．０%

精度に関しては低い数値であった。第一段階と第二段階の差もそれほどないが、やはり第二段階の方が少し劣っている。数値の低い原因は、手法自身の問題以外にも、小説の場面変化の回数がそこまで多くないため、一つ取りこぼすと２０%前後低くなることが考えられる。

５．３　オープンテストと結果

インターネット上に存在する小説サイトから場面変化がある小説をいくつか収集し、テストデータとして、オープンテストを行った。テストデータの選択基準は、会話文が多すぎたり、会話文の中だけで場面変化が行われていたり、回想文などが多かったりしていないかに重点を置いた。場面変化の正解は著者らが手動で作成した。

表３　第一段階オープンテスト結果

	再現率	適合率	F 値
作品①	４７．４%	４８．６%	４８．０%
作品②	８０．０%	２５．０%	３８．１%
作品③	８８．９%	６６．７%	７６．２%
平均値	６２．３%	４９．４%	５５．１%

表４　第二段階オープンテスト結果

	再現率	適合率	F 値
作品①	５０．０%	４５．９%	４７．９%
作品②	８０．０%	２５．０%	３８．１%
作品③	８８．９%	６６．７%	７６．２%
平均値	６４．９%	４８．１%	５５．２%

今回はオープンテストの第二段階の精度が高かった。その理由は、オープンテストの際に使用したデータが童話であるからだと考えられる。童話は子供にも読みやすく書かれたものが多いため、文章が単純になる傾向がある。そのため、パターンが適用しやすかったと考えられる。インターネット小説は作者によって書き方の癖などがあるため、すべてを網羅することは難しいと考えられるが、今回のオープンデータは抽出が上手くいった。

次章では失敗例の分析を通じて手法の問題を考察していく。

６　失敗の分析

精度があまり良くなかったので、抽出が正しく出来ていない原因を詳しく調べた。現在分かっているだけでも大きく分けて以下の四つになる。

・CaboCha の解析ミス
・辞書による失敗
・省略によって正しく抽出ができなかったもの
・パターンの不足により抽出できなかったもの

６．１　CaboCha の解析ミス

長い文章や文章内に――や「」で区切られているものでは CaboCha による解析がいくつか正確ではないものがあった。これは、解析にかける際に「」とその内部を無くしてから解析にかけているからである。また、一文中に並列関係があるものなどの多くは係り受け解析が正しくない。これらは CaboCha によるミスなので、係り受け解析器を変更しない限りでは、解決するのは困難である。

６．２　辞書による失敗

本研究では、日本語語彙大系を辞書作成に使用している。それに関するエラーがいくつかあった。

①　辞書にないもの

辞書に含まれていないものや日本語語彙大系には存在しているが、カテゴリー分けされていないため使用する自作辞書に含めることができない。解決策としては自作辞書にその言葉を手動で追加していく。

例　　「なだれ込む」

②　言い方が古いもの

場所を示すものや動詞の表現が古く、辞書に存在しないもの。解決策としては辞書にその言葉を追加するか、小説のその部分を訂正していく。

例　　「とっつき」　　→　「先」
　　「階（きざはし）」　　→　「階段」

「プラットフォオム」　→　「プラットホーム」

③　削除するもの

完全一致辞書だけでなく正規表現辞書も使用しているため、「～○」などの正規表現のもので場所としてふさわしくないものや間違ったものを抽出してしまう場合があるもの。「○」は任意の文字を表している。

例　「～鼻」半島・岬カテゴリーに存在しているが、顔の「鼻」が場所として抽出されてしまう。
「～中」本来は「～中学校」を指すものである。しかし、「最中」などが場所として抽出されてしまう。

６．３　省略により正しく抽出できなかったもの

主語や動詞、移動先の場所などが省略され、人では簡単に理解できていても、パターンによる抽出が正しくできないものがいくつかあった。

例　私は公園で太郎と遊んでいましたが、夕方になり、太郎は帰ってしまいました。

この例では太郎がどこへ帰ってしまったのかわからない。これは移動先の省略である。

６．４　パターンの不足によって抽出できなかったもの

本研究では、場面変化の抽出の際にパターンマッチング方式を採用している。パターンを増やすことで抽出できるものは増える。しかし、現段階ではこれ以上パターンを増やすことが困難である。むやみにパターンを増やしてしまうと、反例が多くなり、精度が落ちてしまう可能性がある。

上記のように、精度が低い原因には、手法自身の問題点以外にも、既存の基礎ツールの不備や入力文の省略などがあった。次の表は既存の基礎ツールの不備と照応解析に関する主語や場所の省略による不備を省いた結果である。

表５　第一段階クローズドテスト結果（不備除外）

	再現率	適合率	F 値
魚河岸	１００％	１００％	１００％
仙人	１００％	５０．０％	６６．７％
蜘蛛の糸	１００％	４０．０％	５７．１％
十円札	６８．８％	９１．７％	７８．６％
平均値	７８．３％	７２．０％	７５．０％

表６　第二段階クローズドテスト結果（不備除外）

	再現率	適合率	F 値
魚河岸	１００％	１００％	１００％
仙人	６６．７％	３３．３％	４４．４％
蜘蛛の糸	１００％	４０．０％	５７．１％
十円札	６８．８％	９１．７％	７８．６％
平均値	７３．９％	６８．０％	７０．８％

表７　第一段階オープンテスト結果（不備除外）

	再現率	適合率	F 値
作品①	６０．０％	１００％	７５．０％
作品②	１００％	５７．１％	７２．７％
作品③	８８．９％	８８．９％	８８．９％
平均値	７３．１％	８８．４％	８０．０％

表８　第二段階オープンテスト結果（不備除外）

	再現率	適合率	F 値
作品①	６５．４％	９４．４％	７７．３％
作品②	１００％	５７．１％	７２．７％
作品③	８８．９％	８８．９％	８８．９％
平均値	７７．１％	８６．０％	８１．３％

表１～４と表５～８を見比べると、基礎ツールの不備などを除けば、提案手法の有効性が確認できる。すなわち、より精度の高い係り受け解析器と照応解析の技術が加われば精度がよくなると考えられる。パターンの不足も学習データを増やすことで補うことができると考えられる。

７　おわりに

本研究では、小説を読者に理解しやすいように、小説内の場面情報を整理してグラフにすることを目標にしてきた。これにより、読者が登場人物の場面の移動や、場面の移り変わりが理解しやすくなり、そして小説全体の流れがある程度理解しやすくなることが考えられる。

しかし、今回の実験結果では、まだ改善の余地があると思われる。まずは精度を上げることである。精度が低いと読者は小説の場面がつながらず、混乱してしまう可能性がある。既存の研究をうまく利用し、小説内の登場人物の特定も自動でできるようになると提案手法の効率が良くなる。さらに、照応解析などの技術をうまく取り込んで、主語や場所の省略などもうまく抽出できるようになれば、精度も向上する。そして、このシステムを用いて、小説の挿絵を自動生成できるシステムができれば、読者がより理解しやすくなるのではないかと考えている。

参考文献

[1] 馬場こずえ, 藤井敦, 石川徹也：小説テキスト自動分類のためのジャンル推定と人物抽出, 第四回情報科学技術フォーラム講演論文集, pp.67-70 (2005)

[2] 馬場こずえ, 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, 言語処理学会全国大会, D3-3, pp.574-577 (2007)

[3] 高島裕人：小説の登場人物抽出と役割の推定, 東海大学電子情報学部情報メディア学科, 卒業研究 (2005)

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳小倉健太郎, 大山芳史, 林良彦 :日本語語彙大系, 岩波文庫 (1999)