小説における場面変化の情報抽出

04T4027T　斉藤章久

参考にした文章

今回の研究では芥川龍之介の作品である
「魚河岸」「仙人」「蜘蛛の糸」「十円札」の4作品を実験した

作品の選考基準としましては
・著作権が切れえているもの
・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

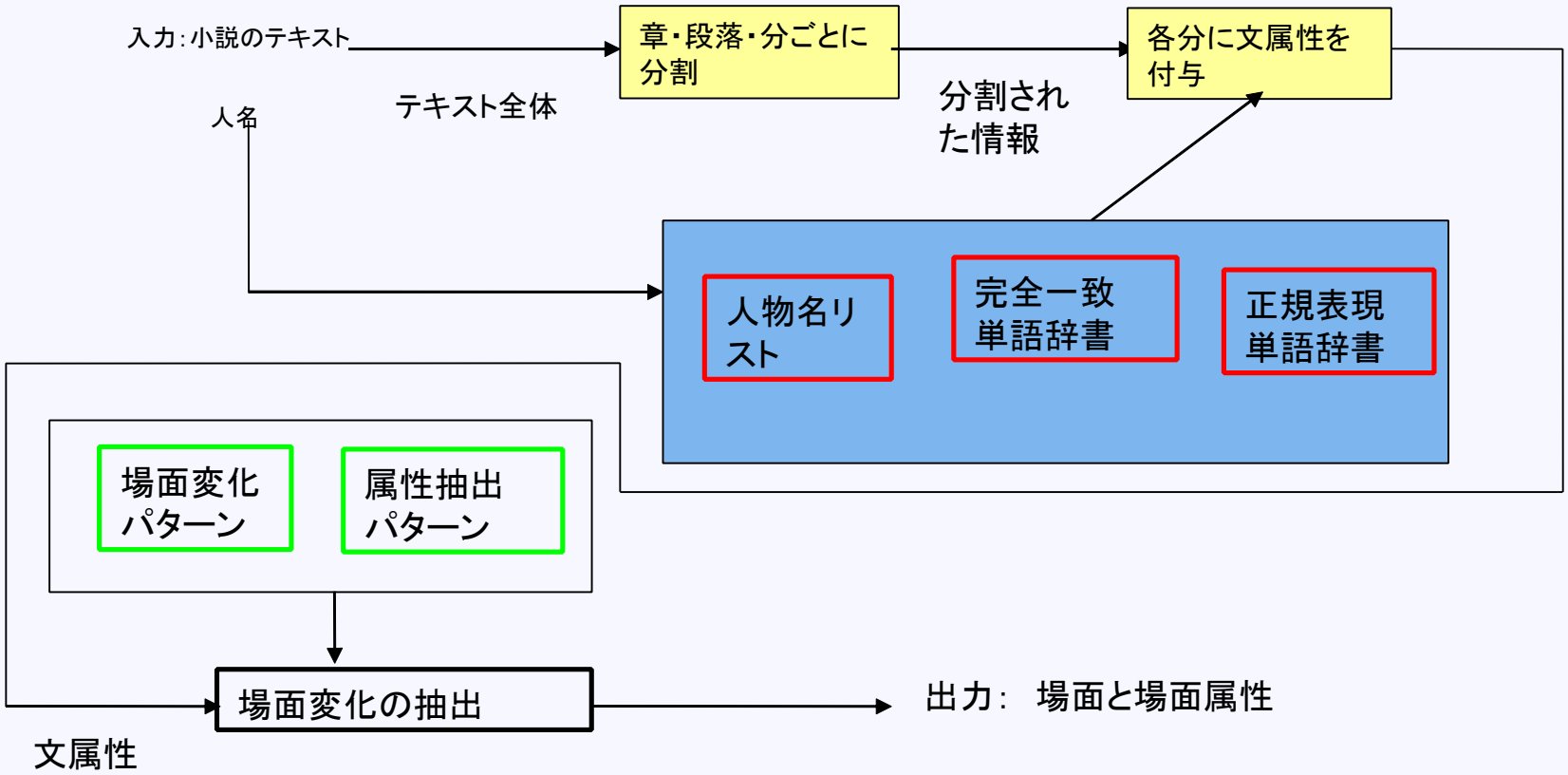
場面変化の定義

本研究で取り扱われる場面変化の定義は

- 場面変化
  場所や状況が切り替わること
  一枚の挿絵で場系を表わす際の単位
- 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や場面自体の様子を表す要素

システムの流れ（1）

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



システムの流れ（2）

図中の「完全一致単語辞書」と「正規表現単語辞書」は日本語語体系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体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。

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。

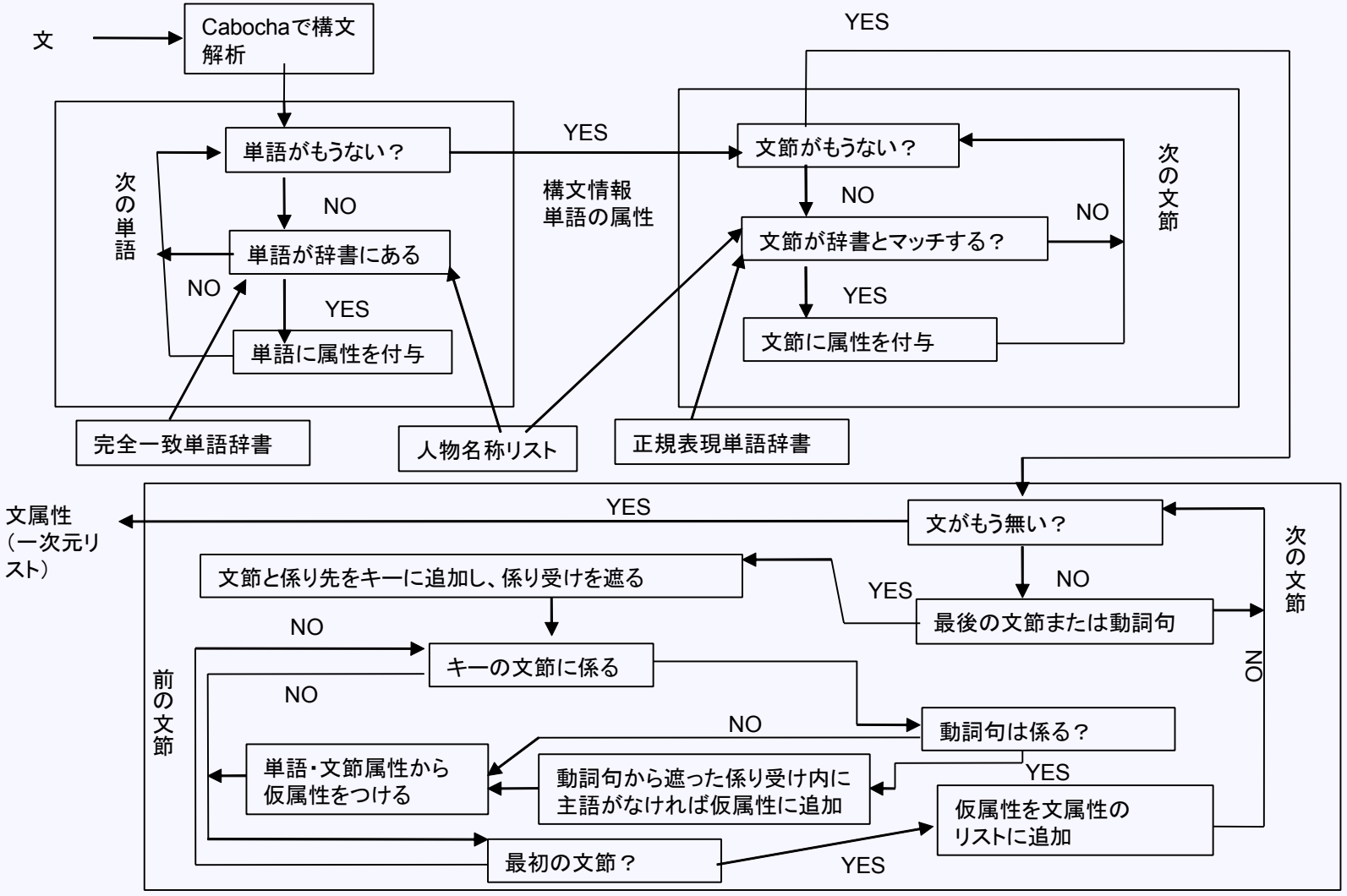
これにより、人物名リストを作成していく。

次に一文ずつ解析をしていくために、章、段落、文ごとに分ける。

それと会話文は考慮しないので会話文の削除も行う

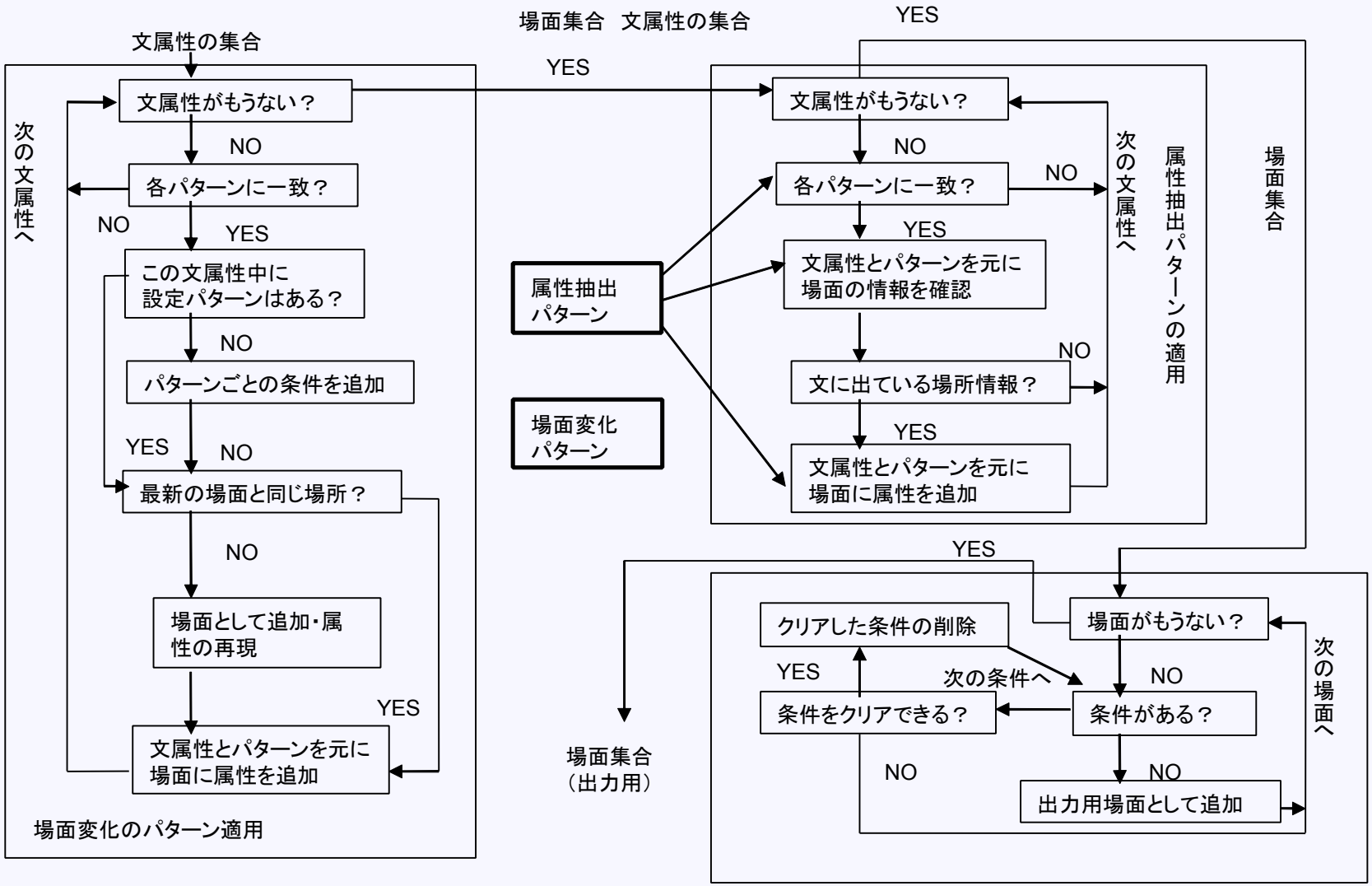
文属性の付与

文章に完全一致単語辞書や正規表現単語辞書などを用いて属性をつける



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変化を抽出する



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。

場面変化パターン（1）

場面変化のパターンとは、入力の一文が場面変化の箇所となっているかどうかをはんだんするためのパターンである。

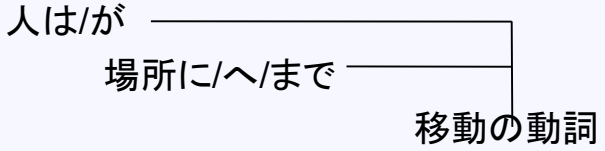
場面変化と場面情報の抽出が、それぞれ前後の文に関係するか否かで4種類にわけられる

①決定パターン・完全
②決定パターン・条件
③候補パターン・完全
④候補パターン・条件

場面変化パターン（2）

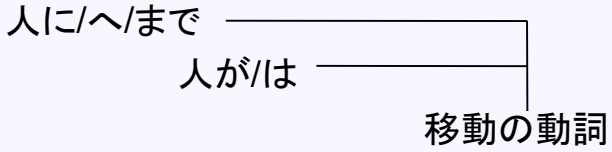
①決定パターン・完全

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

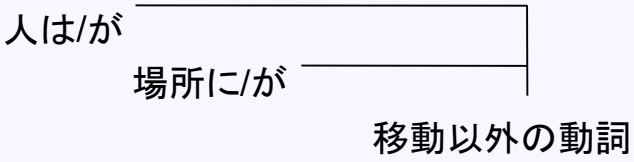
場面変化の抽出に条件はないが、場面属性の抽出には条件があるパターン



場面変化パターン（3）

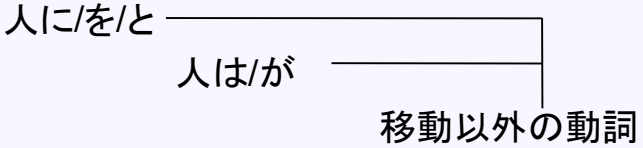
③候補パターン・完全

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



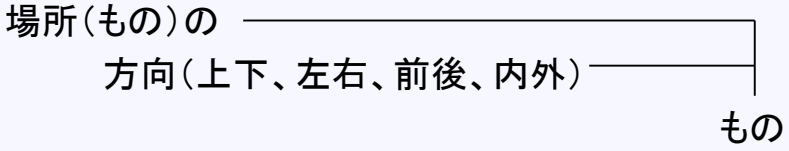
④候補パターン・条件

場面変化、場面変化ともに抽出に条件があるパターン



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。



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行った。

今回の実験はクローズドテストとオープンテストの2種類行い、各実験は第一段階と第二段階に分かれている。

第一段階では場面変化の箇所を特定する。第二段階では場面変化の場所の名前の抽出する。

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回答と比較し、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(%)

実験結果1（1）

第一段階クローズテスト結果

	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%

第二段階クローズテスト結果

	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%

実験結果1(2)

第一段階オープニング結果

	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%

第二段階オープニング結果

	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%

実験１から

・オープニングテストの精度が高かったが、これは用いられたデータが童話であるからと考えられる。

・クローズドテストに関しては精度の低い結果となった。

数値が低い原因は手法自身の問題以外にもあると考えられる

失敗の分析(2)

③省略による抽出ミス

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。

④パターン不足によって抽出できなかったもの

パターンを増やすことによって抽出できるものは増える。しかし現段階で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。

以上の４つが原因である。

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

失敗の分析（1）

精度があまり良くなかった原因は以下の4つになる。

①CaboChaの解析ミス

長い文章や、文章内に「」で区切られているものではCaboChaによる解析が正確ではないものとわかった。「」とその内部を無くしてしまうから解析にかける際にミスをしてしまう。これは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。

②辞書による失敗

辞書にないもの、言い方が古いもので解析ミスしてしまう

また「～中」などで、本来「～中学校」を指すものであるが、「最中」などが場所として抽出されてしまうことがある。

実験結果2(1)

第一段階クローズテスト結果(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%

第二段階クローズテスト結果(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%

実験結果2(2)

第一段階オープニング結果(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%

第二段階オープニング結果(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%

実験2から

- 実験1と実験2の結果を見比べると、基礎ツールなどの不備を除けば、提案手法の有効性が確認できた。

  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、より精度がよくなるであろう

最後に

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。

　しかしまだ改善の余地があると思われる。既存の研究をうまく利用して、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よくなることだろう。さらに照応解析などの技術をうまく抽出できるようになれば精度もますます向上することであろう。