

自然言語テキストからのストーリー抽出と 事象概念構造化システムの構想

大石 顕祐

小方 孝

岩手県立大学大学院ソフトウェア情報学研究科

1. はじめに

Kristeva[1]によれば、テキストは諸種のテキストの相互置換であり、テキスト間相互関連性 (inter-textualite) である。この間テキスト性の理論では、テキストの受容とはその当事者が連想的に他のテキストを記憶から引き出し、現在のテキストと関連付けながら行う経験を意味する。一方制作の側面においては、あるテキストの産出に当って当事者は様々なテキストの記憶や連想の網の中で制作行為を行う。

我々は、物語生成システムを間テキスト性の観点から、入力・分解機構 (既存の物語やその部分を様々な形式の情報に変換)、保存・蓄積機構 (入力され蓄積された情報を加工し、システム内部に保存・蓄積)、再構成機構 (蓄積された情報を基に新たな物語を再構成) から成るものとして再構成した[2]。

これまでの研究では、中嶋ら[3]のシステムが保存・蓄積機構及び再構成機構に相当する。このシステムで入力データは概念表現の一単位が action からなる格フレームとして表現された。この概念表現の作成プロセスは入力・分解機構に相当するが、これまでは手作業で行われていた。

本研究ではこの作業を自動化し、自然言語入力から物語内容を抽出、これを概念構造化するシステムの構成について検討する。このシステムを便宜的に、以下、事象概念構造化システムと呼ぶ。

2. ストーリーの定義

物語論では、物語は大きく分けて物語内容 (story)・物語言説 (discourse)・物語表現の3つの要素から成立しているとされる。物語内容は「物語は何を語っているか」を表す諸要素、すなわち登場人物や舞台などの設定・物語中で起こった出来事とその構成・主題等によって構成される。対して物語言説は視点・時間順序・距離・速度などの修辭的技法や叙法の技法までを含めたものであり、物語表現は小説・映画・音楽などの外的な表現形式を表す。著者はこのモデルから、物語テキストから叙法・修辭の効果を除去し、物語中で確実に起こったと判断できる出来事のみを抽出しそれを時間順に並べ替えることで、物語内容の抽出が可能だと考えた。本研究では、抽出するストーリーを「物語テキストから、物語世界内で実行された出来事を表す文を抽出して、時系列順に並べ替えたもの」と定義する。

3. 事象概念構造化システムの構想

2 節では本研究で抽出するストーリーを定義したが、今回は物語世界内で実行された出来事を表す文を抽出する機能の検討に焦点を絞り、時系列順の並べ替えについては考慮しない。

事象概念構造化システムは、事象概念自動抽出システムと事象概念編集支援システムで構成される。事象概念自動抽出システムは構文解析されたテキストを一文ごとに文法知識と照合す

ることでストーリーを抽出する。一方、同義語の同定や道具格の特定のようなオントロジー知識が必要な部分、出来事の意味的分類のような文間関係の解析が求められる部分については自動化の範疇外として、これらを手動編集ツールである事象概念編集支援システムが行う (図1)。

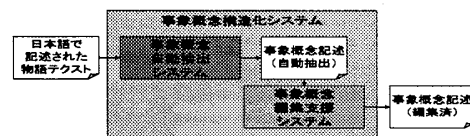


図1 事象概念構造化システムの全体像

以下、事象概念自動抽出システムについて説明する。

3.1. 入力フォーマット

前処理として入力する日本語で記述された物語テキストを文単位で改行、Cabocha で構文解析して、解析結果を京大テキストコーパスフォーマットにしたものを用意しておく。これは、文中での表現に対応する読み仮名、基本形、品詞、活用等が解析された形態素が分節ごとに分かち書きされている。係り受け構造は各文節の先頭に割り振られた番号と係り先の分節番号によって表現されている (図2)。

```
EOS
*0 3D 0/1 1.29866161 セリヌンティウス セリヌンティウス セリヌンティウス 名詞一般 O
は は 助詞-係助詞 O... 記号-読点 O
*1 3D 0/0 3.95832401
深夜 シンヤ 深夜 名詞-副詞可能 O
*2 3D 0/1 0.00000000
王城 オウジョウ 王城 名詞一般 O
に に 助詞-格助詞一般 O
*3 10 0/2 0.00000000 召さ メサ 召す 動詞-自立 五段・サ行 未然形 O
れしれる 動詞-接尾 一段 連用形 O
た た 助動詞 特殊・タ 基本形 O
... 記号-句点 O
EOS
```

図2 入力例

3.2. 出力フォーマット

抽出されたストーリーの個々の要素を事象概念記述とする。これは、行為 (action) を基点に、行為者 (agent)、被行為者 (counter-agent)、行為対象 (object)、行為が行われた場所等を含む格フレームとする (図3)。

```
<doc>
<event id="1">
  <action adverb="深夜">召される
  <agent id="1">セリヌンティウス</agent>
  <counter-agent></counter-agent>
  <instrument> </instrument>
  <object></object>
  <location> </location>
  <goal>王城</goal>
  <from></from>
  <next-location> </next-location>
  <narration></narration>
  <caption></caption>
</action>
</event>
</doc>
```

図3 出力例

On the Extraction of Story from Natural Language and the Structuring of Event Concept

Kensuke Oishi · Iwate Prefectural University

Takashi Ogata · Iwate Prefectural University

3.3. 処理手順

概念表現の自動抽出は 4 の手順で行われる。このうち、①～③までの各機能は action の文中位置を特定し、④～⑥までの各機能は agent、object 等の action 下位要素の文中位置を特定する。①～⑥までの処理は一文単位で行い、③及び⑥の出力結果を⑦が事象概念記述のフォーマットに成型して出力する。

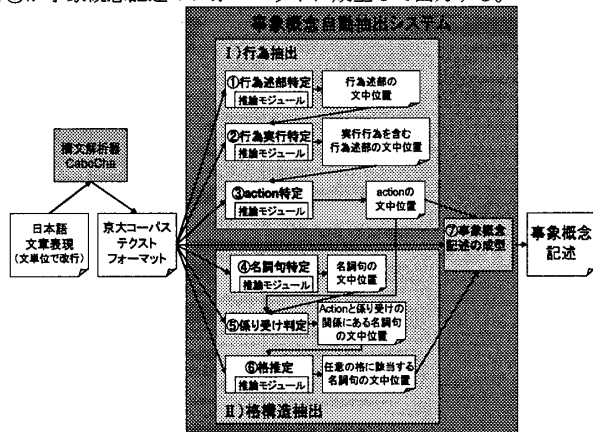


図4 事象概念自動抽出システムの構成

このうち①～④、⑥の各機能は汎用の推論モジュールを呼び出すことにより処理を行う(図5)。

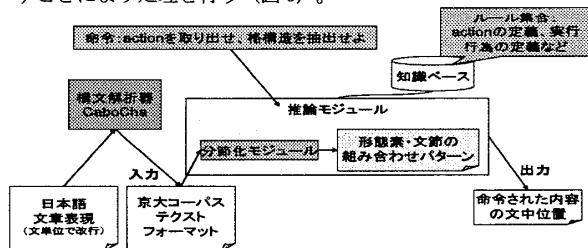


図5 推論モジュールの構成

推論モジュールは、分節化モジュールを呼び出し、入力テキスト中のある一文における任意の形態素組み合わせデータを読み込む。このデータの構成単位は形態素で、文中での表現に対応する読み仮名、基本形、品詞、活用等の情報と、その形態素が含まれる分節の係り受け情報が付随する。このデータを action の定義、実行行為の定義などが記述されたルール集合と照合し、適合した場合、指定された定義に該当する記述の文中位置を示す。ここで用いる知識ベースは文法・語彙知識に基づくもので、呼び出される機能によって異なるルール集合を用いるものとする。

以下に事象概念自動抽出システムの各機能について説明する。

- ① 行為述部特定: 推論モジュールを呼び出し、任意の形態素の組み合わせから行為を表す述部を特定する。ここでいう行為を表す述部には、どのような行為かを現す命題内容に加えて、付随して記述されるボラリティ(命題の真偽)、モダリティ(話し手の心的態度)、テンポラリティ(行為の時間的位置づけ)アスペクチュアリティ(行為の時間的展開の様態)、ボイス(行為の視点)及び敬語・文体(待遇表現)までを含む。また、中核となる命題内容は、表現としては単一の動詞の他、複合動詞、動詞テ形+補助動詞、名詞+補助動詞+動詞化する接辞が接続したもの等の動詞句が考えられる。従って、知識ベースにこれらの定義を記述する必要がある。
- ② 実行行為特定: 推論モジュールを呼び出し、文中のモダリティに着目して①で特定した述部中から物語世界内で実行された出来事を表す行為表現を特定する。ここでいうモダリティとは、言語活動の基本単位としての文の延べ方についての話し手の態度を表し分ける、文レベルの機能・意味的カテゴリー

である[4]。従って、ここでいうモダリティには形態論的カテゴリーであるムードの他、構文的な組み立て形式で表されるものや、叙法副詞による表現を含む。モダリティは大きく分けて意思の表明や行為の要求など文の発話・伝達機能を表すものと、話し手の命題に対する把握の捉え方を表すものがある[5]。前者については、物語世界内で未だ実行されないものとして全て切り捨てる。また後者については、命題内容として書き出された事態の成立・存在を疑いないものとする確認のモダリティ以外の表現を含む文については、その行為が物語世界内で実行されたかどうか不確かだと見做し排除する。

- ③ action 特定: 推論モジュールを呼び出し、②で特定した行為表現から、命題(命題内容・ボラリティ)のみを抽出し、action の値として整形する。
- ④ 名詞句特定: 推論モジュールを呼び出し、任意の形態素の組み合わせが名詞・名詞句であること、及び付随する格助詞を特定する。
- ⑤ 係り受け判定: ③で特定された action と、④で特定された名詞句の間に係り受け関係があるかどうかを、入力テキスト中の係り受け情報から判定する。
- ⑥ 格推定: 推論モジュールを呼び出し、action との係り受け関係が確認された名詞句に付随する格助詞の種類から対応する格を推定し、agent、counter-agent、object、goal、from のいずれかに割り振る。この機能について試作したプログラムでは各名詞句に付随する助詞から格を推定したが、現在構想しているシステムでは、結合価理論[6]に基づき、動詞文型から格を推定することを検討している。
- ⑦ 事象概念記述の成型: ③から入力テキスト中の action の文中位置、⑥から action に係る agent、counter-agent、object、goal、from の文中位置を取得し、事象概念記述のフォーマットに成型して出力する。

なお、今回検討したシステムには含まれていないが、文の時系列順の並べ替えについてはタクシス(taxis)を考慮する必要がある。タクシスとは、1つの出来事と他の出来事との外的時間関係であり[7]、テンポラリティ、アスペクチュアリティの表現を特定することで算出できると考えられる。

4. おわりに

本稿では、自然言語入力から物語内容を抽出し、これを概念構造化するシステムの構成、特に物語世界内で実行された出来事を表す文を抽出する機能について検討した。現在、③および⑥の機能については試作プログラムが完成しているが、これは推論モジュールを用いない不完全なものである。今回構成を検討した事象概念自動抽出システムは現在実装作業中であり、今後評価実験を行う予定である。

参考文献

- [1] Kristeva, J. *le texte du roman*. Mouton Publishers, The Hague, 1970. (谷口 勇 訳, 『テキストとしての小説』, 国文社, 1985.)
- [2] 大石 顕祐・小方 孝・中嶋美由紀・秋元 泰介, 物語生成と間テクスト性の考察, 『ことば工学会 (第 26 回)』, 55-62, 2007.
- [3] 中嶋 美由紀・小方 孝, 物語生成システムと intertextuality-概念の整理と試作の考察-, 『人工知能学会全国大会』, 2E2-2, 2006.
- [4] 宮崎 和人・安達 太郎・野田 春美・高梨 信乃, 『モダリティ (新日本語文法選書 4)』, くろしお出版, 2002.
- [5] 守山 卓郎・仁田 義雄・工藤 浩, 『モダリティ (日本語の文法 3)』, 岩波書店, 2000.
- [6] 小泉 保, 『日本語の格と文法 結合価理論にもとづく新提案』, 大修館書店, 2007.
- [7] 工藤 真由美, 『アスペクト・テンス体系とテキスト - 現代日本語の時間の表現 -』, ひつじ書房, 1995.