1I1-OS1a-3

子供向けテクストを題材とした物語構造の抽出システムの開発

Development of Story Structure Extraction System for Children's Folk Tale Text

宮本 瑠美[*1] Rumi Miyamoto　真部 雄介[*2] Yusuke Manabe　菅原 研次[*2] Kenji Sugawara

[*1] 千葉工業大学大学院情報科学研究科 Graduate School of Information and Computer Science, Chiba Institute of Technology

[*2] 千葉工業大学情報科学部 Faculty of Information and Computer Science, Chiba Institute of Technology

In order to develop the automatic animated movie generation system based on TVML (TV program Markup Language), we are challenging to build ontology for converting a story into an animated movie. Until now we developed 'TVML Choreographer' which can convert a physical posture (3DCG) made by human into TVML. However, the ontology should be a knowledge combining semantic information with TVML. In this paper, we develop a story structure extraction system for children's folk tale text. Our proposed system consists of two phases. The one is to extract keywords (nouns) and verbs. The other phase is to add semantic information, which is dominant conception retrieved from Japanese lexicon dictionary, to keywords text and to convert keywords into symbols. As the result, obtained sequential pattern composed by symbols with concept and verbs is defined as a story structure. This structure is used as a basic knowledge to build ontology for converting a story into an animated movie.

1. 背景

間テクスト性という観点を元に，既存の物語から内容や言説，表現に関する要素を抽出してコンピュータに取り込み，それらを分解・再構成することによって新たな物語を自動で生成する研究が試みられている[小方 2007]. 我々も，そのような研究の文脈において，格フレームで表されたストーリーを入力とし，自動で映像表現を出力する（具体的には映像表現の元となるTVML と呼ばれるマークアップ言語を出力する）システムの開発を目指した取り組みを行っている[真部 2008][真部 2009]. この取り組みの中で重要なものの1つは，ストーリーの格フレーム情報から具体的な映像表現に必要な情報を補完し具体化するためのオントロジーを構築することである. このオントロジーは，構造化された TVML のプリミティブ情報と格フレームが持つような意味情報（語彙の役割のような深層情報）を両方含むものである必要がある. 格フレーム表現を用いたストーリーの概念表現を既存の物語テクストから自動で抽出しようという試みもあるが[大石 2008]，非常に高度な処理や知識が要求されることから，その多くは手作業により作成しており，格フレーム情報そのものを完全な形で自動抽出することは非常に困難であると考えられる.

一方，[赤石 2006]は，テクストから得られる表層的な特徴量を基にして物語構造を抽出・分解・再構成するフレームワークを提案し，有益な情報を取り出す事に成功している. この研究は，単語への深層情報付加などの意味的な処理は一切行わずに名詞のみを取り出し，その連鎖構造に着目してテクストの分節化を実現している点に特徴がある. このことは，格フレームで表現されるような意味的情報以外にもストーリー生成に有用な形式的情報があるということを示唆している.

そこで本研究では，既存の子供向けの物語テクストから表層的な自然言語解析によって容易に得られる情報として，共起に基づく名詞と動詞の組み合わせパターンを抽出し，物語全体の構造を表す手法を提案する. さらに，抽出される物語構造に意味的情報を付加するために，日本語語彙大系辞書[池原 1998]を検索して得られる名詞単語の上位概念情報を付加することで，格フレームのような深層情報を含めた物語構造を得る. この処理によって得られる物語構造は，映像表現や身体運動表現を作成するツールと組み合わせることで，ストーリー自動映像化オントロジーの意味情報として利用できると考えられる.

連絡先：真部雄介，千葉工業大学大学院情報科学研究科，275-0016 千葉県習志野市津田沼 2-17-1, ymanabe@net.it-chiba.ac.jp

2. 提案手法

2.1 処理の概要

図 1 に本研究の目指すシステム全体の処理の流れを示す. 処理は主に 2 段階に分けられる.

第 1 段階では，まず，既存のテクストを入力すると，形態素解析および係り受け解析が行われる. 形態素解析は MeCab[MeCab]，係り受け解析は CaboCha[CaboCha]を利用し

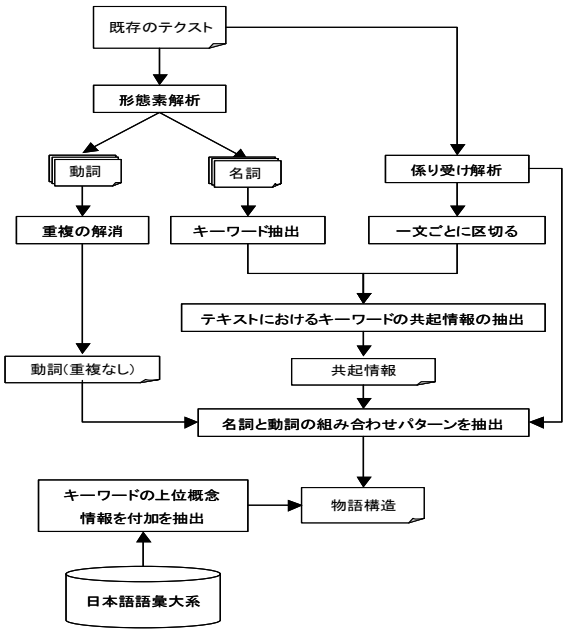

図 1 システムの流れ

ている. 形態素解析では, 動詞(自立)と名詞(一般, 固有)のみを抽出する. 名詞は, 全ての名詞の出現頻度を計算し, 頻度の多いもののみをキーワードとして採用する. また, 動詞は重複するものを除く処理を行う. 入力テクストを一文ごとに区切ったものと抽出したキーワードを使用し, キーワードと動詞の組み合わせパターンを出力する. これが本研究で扱う物語構造の原型である. 係り受け解析結果の利用については後述する.

第 2 段階では, 抽出したキーワードを検索キーとして, 日本語語彙大系辞書を検索して上位概念情報を付加する. また, キーワードはアルファベット記号に置き換え一般化する.

2.2 キーワードと動詞の組み合わせパターン抽出ルール

キーワードと動詞の組み合わせパターンの抽出は, 以下の 3 つの方針に基づいて抽出する.

1. テクストの一文に出現するキーワードの数が 1 つのときは, その文は抽出しない.
2. テクストの一文に出現するキーワードの数が 2 つのときは, その 2 つのキーワードと動詞を抽出する.
3. テクストの一文に出現するキーワードの数が 3 つ以上のときは, 係り受け解析結果を利用する. 係り受け構造をグラフ化したときの各パスにおいて, キーワードが 2 つ以上出現するパスについてキーワードと動詞を抽出する.

頻出度に基づくキーワード抽出のみでは, ノイズとなる不要な語彙を完全に取り去ることができないため, キーワードの共起情報を用いることによって, 本当に必要と思われるもののみを残す.

3. システムの実装と実行結果

システムの開発環境は, OS は Microsoft Windows XP Professional, 使用言語は Microsoft Visual C# 2008 Express Edition を用いた. また, MeCab ライブラリは libmecab.dll を直接インポートして組み込んでいる. CaboCha ライブラリは C#ラッパーである mutterofstar[mutterofstar]を用いている.

図 2 にシステムの実行画面を示す. 今回は, 提案した処理の流れのうち, 第 1 段階までをほぼ実装した. 画面は, 桃太郎[福娘童話集 2010]のテクストを入力した際の実行結果である.

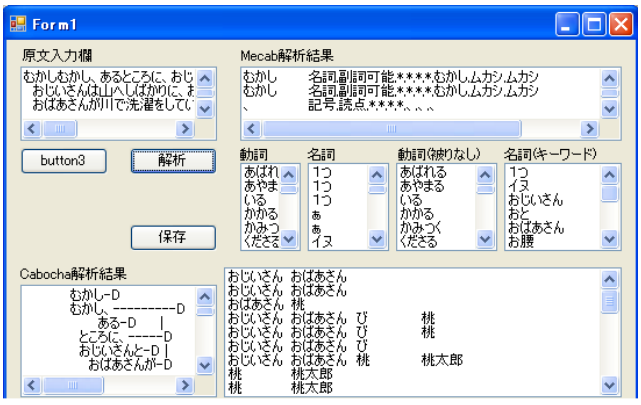

図 2 システムの実行画面

図 3 に, 上記桃太郎を解析して得られた物語構造を示す. ただし, 結果は, システムによって得られた物語構造の原型に対し手作業で第 2 段階の処理(日本語語彙大系を用いた上位概念の付加)を行ったものを示す. 抽出された10個のキーワード A～J の内訳は, A=おじいさん, B=おばあさん, C=桃, D=桃太郎, E=鬼, F=鬼が島, G=イヌ, H=サル, I=キジ, J=宝物である. 物語構造全体を見ると, 桃太郎のストーリーを表す主要な要素と考えられるパターンが抽出できていることがわかる.

```
[A(主体)] + [B(主体)] + 住む
[A(主体)] + [B(主体)] + 行く
[B(主体)] + [C(食料)] + する + 流れる
[B(主体)] + [C(食料)] + ひろいあげる + 持ち帰る
[A(主体)] + [B(主体)] + 食べる + 切る + 飛び出す
[A(主体)] + [B(主体)] + いる
[A(主体)] + [B(主体)] + [D(主体)] + 名付ける
[E(主体)] + [F(地形)] + する + いく
[E(主体)] + [F(地形)] + いく
[G(主体)] + なる + [D(主体)] + もらう
[E(主体)] + [F(地形)] + いく
[E(主体)] + [F(地形)] + いく
[G(主体)] + [H(主体)] + [I(主体)] + 入れた + [D(主体)] + やってくる
[E(主体)] + [J(物品)] + ならべる + ぬすむ
[H(主体)] + [E(主体)] + ひっかく + つつく
[D(主体)] + [G(主体)] + [H(主体)] + [I(主体)] + 帰る
[E(主体)] + 取上げる + [J(物品)] + つむ + 帰る
[A(主体)] + [B(主体)] + 見る + 喜ぶ
```

図 3 桃太郎の物語構造

4. まとめ

本研究では, 名詞と動詞の組み合わせパターンからなる物語構造を抽出するシステムを開発した. 現状では, 日本語語彙大系辞書による語彙の上位概念付加処理は未実装であるので, この処理の実装が課題である. また, システムの全機能を実装した後は, 既に制作済みのシステム[真部 2008][真部 2009]と本システムを統合し, 物語内容を自動映像化するためのオントロジーを構築するシステムへと拡張していく予定である.

参考文献

[赤石 2006] 赤石: 文書群に対する物語構造の動的分解・再構成フレームワーク, 人工知能学会誌, Vol.21, No.5, pp.428-438, 2006.

[CaboCha] CaboCha: Yet Another Japanese Dependency Structure Analyzer, (URL) http://chasen.org/~taku/software/cabocha/

[福娘童話集] 福娘童話集: (URL) http://hukumusume.com/douwa/index.html

[池原 1999] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林: 日本語語彙大系 CD-ROM 版, 岩波書店, 1999.

[真部 2008] 真部, 田中, 長塚, 宮本, 菅原: 物語を自動で映像化するためのツールの試作, 第 17 回文学と認知・コンピュータ II 研究分科会予稿集, 17G-01, 2008.

[真部 2009] 真部: TVML による演技映像生成システムの試作, 第 20 回文学と認知・コンピュータ II 研究分科会予稿集, 20W-01, 2009.

[MeCab] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, (URL) http://mecab.sourceforge.net/

[mutterofstar] mutterofstar (URL) http://download.goo.ne.jp/software/contents/soft/winnt/prog/se446647.html

[小方 2007] 小方: 統合物語生成システム暫定版の諸要素の結合方針, 人工知能学会第 21 回全国大会論文集,1F1-3, 2007.

[大石 2008] 大石, 小方: 物語テクストからのストーリー抽出について, 人工知能学会第 22 回全国大会論文集,1C2-2, 2008.