

令和元年5月27日現在

機関番号：13901

研究種目：挑戦的研究（萌芽）

研究期間：2017～2018

課題番号：17K20028

研究課題名（和文）ブロック玩具をモデルとする日本語文章合成ツールキットの設計と実装

研究課題名（英文）Design and Implementation of a Japanese Text Generation Toolkit Modeled on Interlocking Bricks

研究代表者

佐藤 理史（Sato, Satoshi）

名古屋大学・工学研究科・教授

研究者番号：30205918

交付決定額（研究期間全体）：（直接経費） 4,800,000円

研究成果の概要（和文）：日本語の文章を合成するためのツールキットとして、HaoriBricksを設計・実装した。HaoriBricksは、約1100種類の基本ブロックから構成されており、ユーザーは、それらをどのように組み合わせるかを指定する。指定されたブロックの組み合わせは内部的に保持され、表層化命令によって文の内部構造を經由して最終的に表層文字列に変換される。複数のブロックから構成されるマクロブロックも、基本ブロックと同じように使用できるため、それぞれの応用に対して、ツールの拡張が可能である。

研究成果の学術的意義や社会的意義

日本語の文を合成するためのソフトウェアツールは、これまで存在しなかった。ツールの基盤となる文法体系を定義するとともに、実際に使用に耐えうるツールを実装した点に、本研究の学術的意義がある。潜在的には、文章の自動合成の応用範囲は広い。本研究の社会的意義は、その実現可能性を示した点にある。プログラムによる小説の自動生成の実現は、その一つの応用例である。

研究成果の概要（英文）：In this study, I have designed and implemented a Japanese Text Generation Toolkit, HaoriBricks, which provides 1,100 basic bricks to construct Japanese text. A combination of bricks specified by a user is stored as an internal representation, and it is transformed to the surface string via the sentence structure by the surface realization command. A combination of basic bricks can be defined as a macro brick, which can be used in the same manner as basic bricks. By using this function, we can extend HaoriBricks for a specific application.

研究分野：自然言語処理

キーワード：テキスト生成 日本語生成 自然言語処理 発話表現辞書 発話生成 ブロックモデル

1. 研究開始当初の背景

我々は、2013年より短編小説の自動生成の研究に着手し、**意味の通る1段落以上の日本語の文章を機械的に合成する方法**について研究を進めてきた。2013年度は、既存の文章を切り貼りすることによって文章を合成する方法を検討したが、意味が通る文章を作ることができる可能性が非常に低いことが判明したため、2014年度からは、構造に基づいて文章を合成する方法を採用し、文の内部構造から表層文を生成する文生成器として **Haori** を、文章構造ルールに基づいてテキストプランを作成する **GhostWriter** を設計・実装した。そして、これらのシステムを用いて短編小説を制作し、2015年9月末締切の**第3回星新一賞**に応募した。

日本語の文章の自動生成・自動合成の研究は、散発的な研究を除けば、研究開始時点に至るまでほとんどなかったと言ってもいい。そのため、英語では実現されている **micro planner**（文の構造の決定）や **surface realizer**（文の表層文字列の生成）といった基本的なツールさえ存在しなかった。このため、上記の星新一賞への応募を実現するためには、日本語文章合成のためのシステムをゼロから作り上げる必要があった。

実際に、短編小説の自動生成（文章の自動合成）を実現することを通してわかったことは、次のことである。

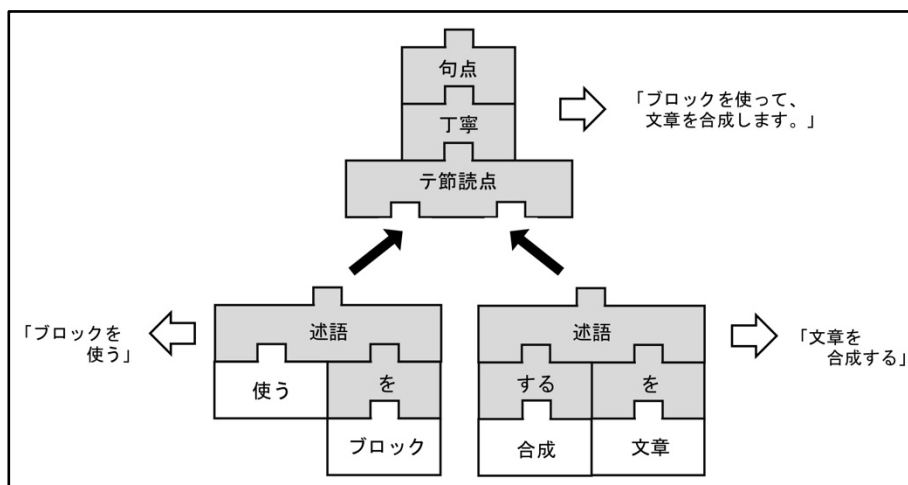
- 文章の自動合成を実現するためには、**大量のテキスト部品**が必要である。ここで、テキスト部品とは、語、句、節、文型パターン、段落パターンなど、文章を構成する、あらゆる大きさのテキストとその断片を意味する。
- テキスト部品は、**適切に抽象化**されていることが望ましい。これにより、**部品のライブラリ化と再利用化が可能**となる。文字列としてのテキスト部品は、文法的な変形が困難なため、再利用できる範囲が限られる。
- テキスト部品を**簡便に組み合わせる方法**が必要である。部品の組み合わせに言語学的な専門知識が必要な場合は、使用できるユーザーが限られる。
- 長編小説の自動生成を実現するためには、テキスト部品の**大規模なライブラリ**が必要である。このような認識より、新たなソフトウェアツールを実現するための本研究を構想するに至った。

2. 研究の目的

本研究では、**基本となるブロックを組み合わせるだけで日本語の文章を自動合成できる、全く新しい方式に基づく文章合成ツールキット**を実装し、その有効性を検証する。このツールがモデルとするのは、レゴに代表されるような**ブロック玩具**である。ブロック玩具では、限られた種類のブロックを組み合わせることによって、多様な作品を構成することが可能である。これと同様に、本ツールキットでは、あらかじめ提供する基本ブロックを組み合わせることによって、多様な文章を簡便に自動合成できるようにする。

具体的には、以下の機能を提供する文章合成ツールを実装する。

- 文および文章を構成するための**基本ブロック**を提供する。これらのブロックは、言語構造（文構造と文章構造）を背後に隠した、**抽象的かつ直感的な部品**とする。
- それらのブロックを**組み合わせる方法**を提供する。すなわち、ブロックを**嵌めるだけ**で、背後では、適切な言語構造が組み上がるようにする。（これを実現するために、それぞれのブロックに、言語構造に関する必要な知識を組み込む。）
- 組み合わせたブロックから、**文法的に正しい文・文章（文字列）を生成する機能**を提供する。
- 組み合わせたブロックを、**マクロブロックとして定義する方法**を提供する。



これらの機能により、以下の目標を達成することを目論む。

- ・ 日本語文法に関する専門知識を持たないユーザーでも、文章合成ツールが利用できること
- ・ マクロブロック定義を利用して、**テキストブロック・ライブラリ**を作成できること
- ・ テキストブロック・ライブラリを利用することにより、文や文章を合成するアプリケーションが容易に作成することができること

つまり、本研究の狙いは、ほとんど機械化されていない**文章合成過程を機械化する手段**を作り上げることである。

3. 研究の方法

解くべき問題は、**文章を組み立てるための基本要素とそれに関わる言語知識をブロックとして抽象化**することである。その中核は、次の3つの問いに答えることに集約される。

Q1: どのような形のブロックを設定するか（どのような組み合わせ方を認めるか）

Q2: どのような基本ブロックを準備すればよいか

Q3: ブロックを導入することによって、どのように抽象化を進めるか

これらの問題は独立ではなく、相互に関連している。次のような方針でアプローチする。

A1 **ブロックの形**: 非常に単純な形状とする。すなわち、すべてのブロックは、1つの突起といくつかのスロットを持つものとする。1つのスロットには、他のブロックの1つの突起を嵌めることができる。

A2 **基本ブロック**: 非常の少数のプリミティブブロックを設定する。それらを組み合わせ、基本ブロックのセットを作成する。

A3 **抽象化**: まずは、抽象度の低いブロックを作成し、段階的により抽象的なブロックをサポートする

4. 研究成果

本研究の成果は、**HaoriBricks** の設計と実装に集約される。

- (1) **HaoriBricks** の基盤となる文法体系(羽織文法)の全体設計を行なうとともに、それに基づく文構造(羽織構造)の設計を行なった。羽織構造は、終端ノードが形態素・語に対応し、非終端ノードが言語構造に対応する木構造である。表層文字列は、この羽織構造から生成される。この生成過程に、活用やテンス付与の処理を組み込み自動化した。これを実現するために、活用体系・接続体系を詳細に定義した。
- (2) **HaoriBrick** によるテキスト生成は、あらかじめ準備されたブロックを組み合わせることで実現される。ユーザーによって指定されたブロックの組み合わせは、ブロック構造として内部的に保持され、表層化命令によって羽織構造に変換され、最終的に表層文字列に変換される。
- (3) **HaoriBricks** は、6種類のプリミティブブロックと、それを組み合わせで定義されるマクロブロックにより構成される。これらを用いて、日本語の基本的な構文構造、および、中核的な付属語・機能表現をマクロブロックとして実装し、Basic パッケージ (ブロックの基本セット) を定義した。Basic パッケージには、約 1100 個(別名も合わせる 2100 個)のブロックが登録されている。このパッケージを用いて、益岡・田窪の『基礎日本語文法-改訂版-』の例文のほとんどを **HaoriBricks** により生成できることを確認した。
- (4) テキストを生成するための **HaoriBricks** のコードをより簡便に記述できるようにするため、羽織構造を変形する機能を導入した。この機能を利用して、述語の取り立て変形を行うブロックを設計・実現した。また、話し言葉の生成を可能にする準備として、「めんどくさい/めんどくせえ」「わるい/わりい」のような話し言葉固有の音変化現象の調査と分類を行なった。
- (5) 組み立てたブロックの内部構造を参照・変形できる汎用的な機構を導入し、これを用いて、多くの総称ブロックを実装した。たとえば、ブロック「丁寧」は、述語が動詞の場合は「ます」を付与し、述語が形容詞の場合は「です」を付与し、述語が形容動詞や判定詞の場合は「だ」を「です」に変換する、のように、対象述語に合わせた変換を実現する。
- (6) 発話文を生成するために、発話表現文書を整備した。この辞書は、74種類の発話意図のそれぞれに対して、使用できる表現文型を列挙した辞書である。この辞書を **HaoriBricks** から利用できるようにした。これにより、発話内容、発話意図、話し方の特徴の3つの入力から、発話文を生成することが可能となり、複数の話し手の発話文のスタイルをコントロールすることが容易となった。
- (7) **HaoriBricks** を利用して、テンプレートに基づく単文生成ツール **HaoriTemplate** を実装した。このツールでは、最小限の付加情報を記述したテンプレート文字列を、内部で **HaoriBricks** コードに変換するため、テンプレートによる生成を行なった後でも、たとえば丁寧体への変換など変形が可能である。複文の作成も、それぞれテンプレートで生成した単文を **HaoriBricks** の機能を用いて組み合わせることで、容易に実現できる。

なお、第5回および第6回星新一賞への応募作品の制作に **HaoriBricks** を使い、第3回の応募時

よりも短時間で短編小説の自動生成を実現できることを確認した。

5. 主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 3 件）

- ① 木村遼, 夏目和子, 佐藤理史, 松崎拓也. 発話表現文型辞書を利用した多様な発話文生成機構. 2018 年度人工知能学会全国大会(第 32 回), 2018.
- ② 佐藤理史. 述語取り立て形式の整理と文生成器への実装. 言語処理学会第 24 回年次大会, 2018.
- ③ 宮崎千明, 佐藤理史. テキストへのキャラクタ性付与のための音変化現象の分類. 言語処理学会第 24 回年次大会, 2018.

〔その他〕

- ① 佐藤理史. 「コンピュータが小説を書く日」はまだ来ない. AJALT2018, pp. 22-27, 2018.
- ② AI は星新一、ユーミンを超えるか. 論点スペシャル, 毎日新聞, 2018 年 1 月 5 日朝刊.
- ③ 佐藤理史. コンピュータが小説を書く日. FM FESTIVAL 未来授業. TOKYO FM, 2017 年 11 月 3 日放送.
- ④ 人工知能 芸術家になれるか. 北海道新聞, 2017 年 4 月 12 日朝刊.

6. 研究組織

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。