科研費

科学研究費助成事業研究成果報告書

令和 3 年 5 月 3 1 日現在

機関番号: 13901

研究種目: 基盤研究(B)(一般)

研究期間: 2018~2020

課題番号: 18H03285

研究課題名(和文)日本語文章の構造モデルとその段階的詳細化による文章自動生成機構

研究課題名 (英文) Structural modeling and automatic generation of Japanese text

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) 13,300,000円

研究成果の概要(和文):本研究では、文章の背後に構造が存在することを仮定して、文章生成技術および文章解析技術の実現に取り組んだ。生成側では、文合成ツールHaoriBricks3、文型集ツールMicroGenerator21を実装した。これらを利用する文章生成ツールGhostWriter21は、複数回の改訂を経て文章の決定稿を作成する人間の文章作成過程を模倣した新しいモデルに基づいている。これらを用いて、テレビの60秒広告のシナリオを自動生成するシステムを実現した。解析側では、文章構造を把握する足掛かりとして、日本語文の文末述語を同定・解析するツールPanzerを実装し、文間関係の把握に有効に働くことを確認した。

研究成果の学術的意義や社会的意義本研究は、これまでの言語処理が主に対象にしてきた「文」ではなく、複数の文から構成される「文章」を対象としている。文章の背後には書き手の思考があり、読み手を説得する構造がある。文章を対象とする自動処理では、この構造をとらえない限り、本質的なことは実現できない。本研究では、数百字の文章の自動生成のために必要なツール群の構成を明らかにし、それらを実装した点に学術的意義がある。これらのツールのうち、文章生成ツールは、文章の作成過程を改訂処理の繰り返しとして捉える新しいモデルに基づいている。このモデルは、文章作成をコンピュータと人間の共同作業化する新しい可能性を内在している。

研究成果の概要(英文): In this research, we have implemented generation and analysis technologies for Japanese text of multiple sentences. On the generation side, we have implemented the Japanese sentence synthesis tool HaoriBricks3 and the sentence-pattern collection tool MicroGenerator21. For text generation, we have implemented a system GhostWriter21, which is based on a new model that imitates the human text writing process of producing a final version through multiple revisions. Using these systems, we have realized a system that automatically generates scenarios for 60-second advertisements on TV. On the analysis side, we implemented Panzer, a tool for identifying and analyzing sentence-final predicates in Japanese sentences, as a foothold for grasping text structure. We have confirmed that it works effectively for understanding inter-sentence relations.

研究分野: 自然言語処理

キーワード: 自然言語処理 文章生成 文章解析 広告の自動生成

1.研究開始当初の背景

自然言語処理技術は、文単体の解析技術に関しては一定の成果を示してきた。しかしながら、複数の文から構成される「文章」の読解や生成などの、真に現実的な問題に対しては、多くの難問を抱えている。我々のグループは大学入試問題の国語「現代文」の読解問題を解くシステムの研究を 2013 年より行ったが、記述式の読解問題に対しては、まったく歯が立たなかった。それには多くの原因があるが、最大の原因は、多数の文から構成される文章に対して、それをどのように捉えればよいかという適切なモデルが存在しないことにある。

すべての文章は、誰かによって書かれたものである。それは、なんらかの目的のために計画され、肉付けされ、最終的に文字の1次元の並びとして表層化されたものである。すなわち、文章の背後には、書き手の思考の構造があり、読み手を説得する構造がある。これは、文の構文構造とはまったくの別物であり、いわば、思考や説得の典型的な型に基づく構造である。文章の自動生成や自動解析を実現するためには、この構造を適切にモデル化するとともに、その構造と(表層テキストに現れる)言語表現との関係を明らかにする必要がある。

2.研究の目的

本研究では、「文章をどのようにモデル化すれば、文章の自動生成が容易になるか」を中核的問いに据え、文章生成のための文章の構造モデルの確立と、そのモデルに基づく文章生成機構の実現に取り組む。大学入試の評論読解問題の自動解答を目指す研究で明らかになりつつある文章の構造的把握に関する知見と、短編小説の自動生成の研究で培ってきた文生成・文章生成技術をより統合的に発展させ、複数のレイヤを持つ文章の構造モデル(テキストプラン)と、その段階的詳細化によって文章を生成する機構の実現を目指す。

3.研究の方法

研究は、通販商品のテレビ広告のシナリオの自動生成の実現を中核とする生成系からのアプローチと、評論読解問題を解くための文章構造の把握を目的とした解析系からのアプローチの両側面から進める。わかりやすい最終目標として、テレビ広告のシナリオの自動生成システムの実現を掲げる。

本研究では、アプリケーションを中心としたアプローチを採用する。すなわち、設定したアプリケーションを実現するための技術開発を行い、得られた方法と知見をツール(基盤ソフトウェアシステム)化することを目指す。

4.研究成果

(1) 生成系からのアプローチ

日本語文合成ソフトウェア Haor i Brick3 の実現

文章を合成するためには、その前段階として文を合成できなければならない。これを実現するソフトウェアツールとして、HaoriBrick3 (HB3)を実装した。HB3 では、ブリックコードと呼ぶRuby コードで、どのような日本語文を合成するかを記述する。このブリックコードを評価すると、ブリック構造および羽織構造という内部構造を経由して、最終的に表層文字列が生成される。HB3 は、文生成に必要な文法事項を整理・再構成した羽織文法に基づいている。羽織文法は、「日本語の文は、機能語によって組み立てられており、文法(文の成り立ちとその規則性)を明らかにすることは、機能語を列挙し、その振る舞いを記述することである」という基本思想に基づいている。HB3 の内部には、約 600 語の機能語の振る舞い(活用型と接続型)が厳格に記述されており、これらと数種類の構造によって、文構造が規定される。HB3 は、約 2000 種類の部品(ブリックと呼ばれる)を提供しており、これらの部品の組み合わせ方法をブリックコードとして記述する。ブリックコードは、Ruby コード(プログラム)なので、任意の部分を変数化したり、制御構造を埋め込むことが可能である。この自由度のおかげで、HB3 を用いると、汎用性の高い文合成コードを記述することができる。HB3 は、本プロジェクトの広告シナリオ自動生成で使用された以外にも、複数の異なるアプリケーションで使用された。

改訂を中心とした文章作成モデルの提案

これまでの文章生成のモデルの基本的な考え方は、「意味表現を表層テキストに変換(翻訳)する」というものであったが、このモデルでは、人間の文章作成プロセスの一部しかモデル化できていない。そのため、複数回の改訂を経て決定稿を作成する過程をモデル化した新しい文章作成のモデル(文章改訂モデル)を考案した。このモデルでは、テキストの改訂処理がその中核と

なる。改訂処理の入力は、現在のテキストと伝えたい情報の2つであり、出力は、改訂されたテキストとそのテキストで伝わる情報の2つである。伝えるべき情報をテキスト化するのではなく、テキスト化の結果として伝わる情報が定まる点、および、テキスト化の能力が完全ではない(伝えたい情報を完全に伝えるテキストを必ずしも作れるわけではない)とする点が、従来のモデルとは決定的に異なる。

広告シナリオの生成支援システムの実現

通信販売商品のテレビ 60 秒広告のシナリオの自動生成機構を持った作成支援システムを実現した。このシステムは、3年間に渡って何度か作り直したが、最終的なシステムは、上記の文章改訂モデルを基本モデルとしている。ユーザーは、作成する文章の仕様(盛り込みたい内容)をシステムに指示する。システムは、この仕様をできるだけ満たすテキストと、そのテキストを完全に記述する(1対1に対応する)仕様を出力する。ユーザーは、出力されたテキストに満足できなければ、仕様を修正してテキストを再生成(改訂)することができる。これを繰り返すことにより、テキストを望ましい形に近づけていく。

本システムは、文合成ツール Haori Bricks3、文型集ツール MicroGenerator21 (MiG21)、文章 生成ツール GhostWriter21 (GW21)の3つのツールを使って実現されており、コンテンツ(文章 生成に必要な知識群)としては、広告シナリオでよく使われる文型集、広告シナリオの典型的な文章構造集、および、広告シナリオの実例集が含まれている。本システムを実用レベルに持っていくためには、コンテンツが不足しており、その充実が課題として残った。

(2) 解析系からのアプローチ

対比構造の分析と検出

文章に含まれる重要な構造の一つに対比構造がある。入試問題の評論文を対象に、対比構造の 具体例を分析し、対比構造を検出する方法について検討した。検出方法は3ステップで構成され る。まず、対比構造を構成する2つの対比要素の転換点となる転換フレーズを検出する。次に、 そのフレーズの出現位置に基づいて、それぞれの対比要素の範囲を推定する。その後、それぞれ の対比要素をコンパクトに要約する対比キーワードを抽出する。実際に、過去の入試問題の評論 文を対象に、一文内の対比構造、および、一段落内の対比構造に対して検出実験を行った。前者 では転換フレーズの検出が、後者では対比要素範囲の推定が難しいことが明らかになり、対比構 造に関連する入試問題を正しく解くことは、現時点では困難であることがわかった。

文末述語解析システム Panzer の実現と文間接続関係の推定

文章構造の把握のためには、それぞれの文のタイプ(文章中における役割)の把握と、文間接続関係の把握が不可欠である。日本語の文では、その文がどのようなタイプの文であるかの情報は、主に文末の述語部分に現れる。そこで、日本語文の文末の述語の範囲を同定し、それがどのような要素から構成されているかを分析する日本語文末解析器 Panzer を実装した。日本語の述語は、おおよそ、核となる内容語(動詞・形容詞・名詞)に、接尾辞や助動詞等の付属要素が接続することによって構成されている。Panzer ではこれらの付属要素を厳密に定義するとともに、211種類(432 エントリ)の述語複合辞(助動詞的に働く連語)を定義し、付属要素として認識できるようにした。なお、Panzer の出力は Haori Bricks3 で表層文に復元可能であるため、復元テストによる解析エラー検出が容易である。

Panzer による分割解析結果を利用して、文末述語を用いた2文間の接続関係推定を実装した。 具体的には、入試で出題される接続詞補充問題(空欄を埋める適切な接続詞を選択する)を対象 として、前後2文の文末述語の情報のみからどの程度正しく接続詞を推定できるかを調べた。既 存の解析器を用いた場合と比較して、Panzer の解析結果を用いると、統計的に優位な性能向上 が見られた。

大学入試問題を題材とした文章の理解

大学入試問題を題材として、「問題を解くことができる=問題文(文章)を理解した」という考え方に立って、化学の熱化学計算問題、および、数学の確率問題を対象に、文章の理解の一端を明らかにした。入試問題の問題文には、対象に固有な典型的な文章構造が存在し、それらの構造を構成する文を認定することはそれほど難しくない。しかしながら、問題を解くために必要な情報を引き出すには、言語表現のみならず、領域知識や常識が必要となる。熱化学計算問題では、物質の化学式や物質量、熱化学反応の種類とそれに対応する熱化学方程式の一般式等の知識なしには、問題を解くことができない。数学の確率問題では、確率の概念を構成するための基本的な概念(試行、母集団、事象)の理解が不可欠である。「文章を読む(理解する)」とは、単に表層上の意味を把握することではない。文章から引き出した情報と読み手が持つ知識とのインタラクションに、理解の本質がある。

(3) 全体のまとめ

文章の構造モデルという点では、本研究ではそれほど大きな進展は得られなかった。文章の構造は、対象とする文章の種別に大きく依存する。そのため、汎用モデルは、たとえば、起承転結のように非常に抽象的なレベルの記述となり、実際の文章生成・解析処理にはほとんど役に立たない。つまり、文章生成・解析のために、それぞれの種別に対して構造モデルを設定することが必要となる。

文章生成技術については、文合成ツール、文型集ツール、文章生成ツールの3つルールで、文章生成システムを構築することができることが明らかになった。文章生成ツールでは、文を部品として、文章を組み立てる方法を記述する。部品となる文は、文型集より提供される。それぞれの文型は文合成ツールによって記述され、1つの文型で多様なバリエーションを生成できる。これら3つのツールが担当すべき範囲がはっきりしているため、以前に比べて、必要な作業が明確となり、文章生成システムの実現が容易となった。

文章解析技術は、汎用的なツール化が難しい。Panzer は、文章の構造把握に必要な足掛かりを提供するが、文章ではなく文の解析ツールである。文章解析の最大の障害は、その最終ゴールをどこに設定するかが明確ではないという点にある。そのため、研究の後半では、問題を解くという明確なゴールを設定して、問題文(文章)の解析に取り組んだ。そこでは、領域知識をうまく使うことが重要であり、文章の理解とは、文章から引き出した情報と読み手が持つ知識とのインタラクションにあるという考えに至ることになった。

以上をまとめると、本研究の最大の成果は、文章生成のための3つのツールの実現による広告 シナリオの生成支援システムの実現である。

5 . 主な発表論文等

「雑誌論文】 計2件(うち査読付論文 2件/うち国際共著 0件/うちオープンアクセス 2件)

【雜誌論文】 計2件(つら直読的論文 2件/つら国際共者 0件/つらオープファクセス 2件)		
1 . 著者名	4.巻	
宮崎千明,佐藤理史	26	
2.論文標題	5.発行年	
発話テキストへのキャラクタ性付与のための音変換表現の分類	2019年	
, , , , , , , , , , , , , , , , , , ,		
3.雑誌名	6.最初と最後の頁	
自然言語処理	407-440	
	107 110	
掲載論文のDOI(デジタルオブジェクト識別子)	査読の有無	
10.5715/jnlp.26.407	有	
10.107.107.107		
オープンアクセス	国際共著	
オープンアクセスとしている(また、その予定である)	-	
3 777 27(20 27) (27) (27) (27)	L	
1.著者名	4 . 巻	
·····································	27	
江際生义	21	
2 . 論文標題	5 . 発行年	
	1 - 1,- 1	
HaoriBricks3: 日本語文を合成するためのドメイン特化言語	2020年	
0 185+47	- 日初に目後の五	
3.雑誌名	6.最初と最後の頁	

411-444

査読の有無

国際共著

有

〔学会発表〕 計10件(うち招待講演 3件/うち国際学会 1件)

オープンアクセスとしている(また、その予定である)

1.発表者名

自然言語処理

オープンアクセス

夏目和子, 佐藤理史

10.5715/jnlp.27.411

2 . 発表標題

発話文表現文型辞書の設計と編纂

掲載論文のDOI(デジタルオブジェクト識別子)

3.学会等名

言語資源活用ワークショップ2019

4 . 発表年

2019年

1.発表者名

今頭 伸嘉, 平良 裕汰朗, 宮田 玲, 佐藤 理史

2 . 発表標題

通販商品を対象としたテレビ60秒広告シナリオの構造分析と自動生成

3 . 学会等名

情報処理学会研究報告, Vol.2019-NL-243 No.16

4.発表年

2019年

1.発表者名
生,光秋有有 佐野正裕,佐藤理史,宮田玲
2.発表標題
文末述語における機能表現検出と文間接続関係推定への応用
3 . 学会等名
言語処理学会第26回年次大会,B6-3
4.発表年
2020年
1.発表者名 佐藤理史
<u>性脉连头</u>
2.発表標題
文章生成研究は楽しい
3.学会等名
情報処理学会第239回自然言語処理研究会(招待講演)
2019年
1. 発表者名
佐藤理史
2 . 発表標題 コンピュータが小説を書く日
3.学会等名
AIと日本語教育 国際学術研究会(招待講演)(国際学会)
4.発表年 2019年
2010
1 . 発表者名
佐野 正裕, 佐藤 理史, 松崎 拓也
2.発表標題
評論文における対比構造とその自動検出
3.学会等名
3.字云寺石 言語処理学会 第25回年次大会
4 . 発表年
2019年

1. 発表者名 平良 裕汰朗, 佐藤 理史, 宮田 玲, 今頭 伸嘉
2 . 発表標題 ダイレクト広告コピー文の分析と自動生成
3.学会等名言語処理学会第25回年次大会
4 . 発表年 2019年
1.発表者名 佐藤理史
2.発表標題 動く文法!? -文をブロックで組み立てる-
3 . 学会等名 日本語文法学会第21回大会 シンポジウム(招待講演)
4 . 発表年 2020年
1.発表者名 経種直之,佐藤理史,小川浩平,宮田玲
2 . 発表標題 熱化学計算問題の段階的解釈に基づく熱化学方程式の立式
3.学会等名 言語処理学会第27回年次大会
4 . 発表年 2021年
1.発表者名 岩間純輝,佐藤理史,小川浩平,宮田玲
2 . 発表標題 数学・確率問題を対象とした条件記述の自動解釈. 言語処理学会第27回年次大会
3.学会等名 言語処理学会第27回年次大会
4 . 発表年 2021年

〔図書〕 計1件

COO HILL	
1 . 著者名 新井紀子, 東中竜一郎 編 佐藤理史ら (総勢26名)著	4 . 発行年 2018年
2.出版社 東京大学出版会	5.総ページ数 288
3 . 書名 人工知能プロジェクット「ロボットは東大に入れるか」	

〔産業財産権〕

〔その他〕

_

6.研究組織

 •	W1フしか上が40		
	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考

7.科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------