様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　２９　日現在

機関番号：　１３９０１
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７４８

研究課題名（和文）文章の読解と産出のための言語処理技術

研究課題名（英文）Language Processing Technologies for Understanding and Production of 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）　　12,200,000円

研究成果の概要（和文）：文章の読解では、センター試験形式の評論読解問題を対象に、多くの特徴量を利用した二段階選抜法を用いたソルバーを実現した。このソルバーは、対象とする問題の約半分を正しく解くことができた。文章の算出では、日本語の文生成器HaoriBricks、および、文章生成器GhostWriterを実装し、シナリオ文法に基づく文章生成を実現した。これらのシステムを用いて自動生成した短編小説を星新一賞に応募した。

研究成果の概要（英文）：This study contains two sides of language processing. On the understanding side, we have developed a solver for essay comprehension questions of the National Center Test, which are multiple-choice questions. The solver determines an answer by two-step screening; first it selects two choices among five, then it selects the best between them. Both selection steps are trained on examples. On the production side, we have implemented a Japanese sentence generator, HaoriBricks, and a Japanese text generator, GhostWriter. By using these systems, we produced several short stories, and submitted them to a Japanese literary award, Hoshi Shinichi Award.

研究分野：　自然言語処理、人工知能

キーワード：　文章の読解　文章の産出　文章生成　節境界　評論読解問題　短編小説の自動生成

１．研究開始当初の背景

現在の自然言語処理の中核技術は、機械翻訳への応用を暗に想定した「文の構文解析技術」と、情報検索のインデキシング（索引語付与）に端を発する「語のリストを中心とした技術」に大別できる。これらの技術は、機械学習の援用により、1990 年代後半より大きな進展を見た。

文の構文解析は、独立した文に対して理想的な解析結果（構文解析木）を定義し、それと同一の構文解析木を出力することに注力する。その処理は、文の構造を明らかにするが、文がひとまとまりの文章の一部である事実は全く考慮されない。文間にまたがる処理は、少数の個別問題（たとえば、指示語・代名詞等の参照先の同定）が切り出されて研究されているが、ひとまとまりの文章をどのように解析すればよいか、あるいは、その結果をどのように表現すればよいかという問題は、真剣には取り組まれていない。

一方、情報検索を支える基盤技術の一つである、文章に対する索引語付与は、処理対象（入力）がひとまとまりの文章であるという点で、文解析を補完するように見える。しかしながら、処理結果（出力）は内容を表す少数の語のリストであり、文解析と接点を持ち得ない。

これら 2 つの技術は、ある意味で 2 つの極を形成するが、同じ欠陥を持つ。すなわち、構文解析は、文の構造に着目するが、文章の構造は無視する。もう一方の索引語付与は、文章を処理対象とするが、文章が内在する構造をすべて無視する。

コンピュータによる文章の読解や産出に取り組もうとすると、これらの既存技術だけでは、ほとんど手も足も出ないことを思い知らさせる。たとえば、センター試験『国語』の評論読解問題では、3500 字強の本文に対して設問が提示されるが、本文の各文の構文構造が正しく得られたとしても、設問に正しく解答できることには直接つながらない。なぜなら、設問で問われるのは、文章の論理構造（文間関係や段落間関係に基づく論旨展開）であり、それぞれの文の個別の伝達内容ではないからである。文章生成に至っては、もっと悲惨な状況であり、どのようにアプローチすればよいかさえ明確ではない。少数のキーワードのリストから、意味のあるひとまとまりの文章を生成できないことは、明白である。

我々は、2013 年度に、2 つの意図を持って、大学入試『国語』の読解問題の自動解法と、短編小説の自動創作の研究をほぼ同時にスタートさせた。1 つ目の意図は、これまで自然言語処理が視野に入れてこなかった（あるいは、無理だと考えられてきた）新たな問題に挑戦することにより、技術の再編と新規開発を強力に押し進めようという意図である。もう 1 つは、解析系（読解）と生成系（産出）の両方を同時に考えようという意図である。文章は、書き手により産出され、読み手により読解される。その過程で情報が伝達される。言語（文章）を媒体とした情報伝達をモデル化するには、書き手と読み手という 2 人のプレーヤー、産出と読解という 2 つの側面を考慮することが不可欠である。

２．研究の目的

本研究の目的は、コンピュータによる文章の読解（解析）および産出（生成）の実現に向けて、言語処理技術の再編と新規開発を行なうことにある。これまでの言語処理技術の中心は、機械翻訳を暗に想定した「文の解析技術」であったが、大学入試『国語』読解問題への挑戦により、それだけでは文章を読み解くことができないことが明らかになった。独立の文ではなく、ひとまとまりの文章をどのように解析すべきか、あるいは、ひとまとまりの文章をどのように生成しうるか。本研究では、「節」という言語単位を処理の中核要素に据え、解析系と生成系の両者を視野に入れて言語処理技術を再構築し、大学入試『国語』読解問題の自動解法と短編小説の自動創作の実現に挑戦する。

３．研究の方法

(1) 節の分析・定式化と節境界認定ツールの実現：現代日本語書き言葉均衡コーパスのコアデータを対象に、節の分析・定式化と節境界認定ツールの開発を同時並行的に行う。自動認定結果に対してエラー分析を行うことにより、認定ルールの精錬と節形式の分類・定式化を推し進める。
(2) 大学入試問題の評論読解問題を題材とした文章読解技術：文章読解では、その出力（読解結果とは何か）が明確ではない。そこで、入試問題の読解問題を解くプログラム（ソルバー）を実現することを目標に、どのような方法で人間並みの性能が得られるかを探求する。同時に、読解問題を解くために文章から抽出すべき情報を明確にし、それらの情報を抽出するための技術についても探求する。
(3) 短編小説の自動生成の実現と、そのための文章生成技術：星新一賞への応募を目標に、コンピュータプログラムで自動生成したと一目で見破られないような短編小説を自動生成する技術を探求する。日本語を対象に、文を自動合成するシステム、および、文章を自動合成するシステムをどのような形で実現すべきかを検討し、その実現に取り組む。

４．研究成果

(1) 節境界認定ツール Rainbow の実現と節の定式化：2015 年度に境界認定ツール Rainbow を再実装するとともに、現代日本語書き言葉均衡コーパスのコアデータの一部（４つのレ

ジスタ）に節境界を付与し、その結果を公開した。この過程において、日本語の節形式の整理がかなり進み、残された問題がどこにあるかが明確となった。なお、この研究は、国立国語研究所の丸山岳彦氏と協力して行った。

(2) 大学入試評論読解問題ソルバーの深化：2015 年度には複数の特徴を利用したセンター試験『国語』評論読解問題用のソルバーを実現し、模試を含むセンター型の選択式問題の 4 割強が解けるレベルに到達した。2015 年秋に実施した、進研模試を利用した公開性能評価（東ロボフォーマルラン）では、対象とする 8 問中 5 問に正解するという好成績を収めた。2016 年度には、このソルバーをさらに改良した。具体的には、語の重要度を加味した特徴を新たに導入するとともに、５択から正解を選ぶ過程を１段階から２段階（まず上位２位までを定め、最後にもう一度どちらを選ぶか決める）に変更した。これらの改良により、性能は統計的に有意に向上した。2016 年秋に実施した公開性能評価では、評論読解問題８問中６問に正解した。これらの一連の研究成果により、ソルバーの能力は、平均的な高校生と同等レベルに達した。

(3) 手がかり表現に基づく文の内容および文間関係推定：記述式の評論読解を解くための文章解析の最初のステップとして、内容属性と関係属性と呼ぶ２種類の情報の付与に取り組んだ。文章を構成するユニット（おおよそ文または節)の内容属性とは、そのユニットの文章中における役割 （たとえば、著者主張や提起など）である。一方、関係属性とは、２つのユニット間に対する関係（たとえば、順接、逆接、転換など)である。大学入試問題の本文として出題される評論文に特徴的な内容･関係属性を 26 種類に分類し、それらを示唆する言語表現と対応づけて整理した（これらの言語表現の中には、内容属性と関係属性の両方を示唆する表現も存在する)。この整理に基づいて、テキスト中に含まれるこれらの言語表現を認定するシステムを実装し、内容属性と関係属性の自動付与を実現した。

(4) 文生成ツールの実現：2015 年度は、それまで開発してきた文生成ツール Haori に、述語文節の形式変換機能と節挿入機能を追加し、複文を合成できるように機能拡張した。これにより、複数の単文から複文を合成することが可能となった。Haori を文章生成（ショートショート生成）に実際に使用し、所望のテキストを出力できることを確認したが、より抽象的な入力からの文生成能力を持つことが望ましいことも判明した。これを受けて、2016 年度に Haori を抜本的に見直し、より簡便に文や文章を自動合成できる HaoriBricks を設計・試作した。HaoriBricks は、ブロック玩具のように、小さな部品を組み合わせてより大きな部品を作ることが可能であり、この機能を用いて、単語から文、文から文章を組み上げることができる。さらに、HaoriBricks はプログラミング言語 Ruby のライブラリとして実現されているため、Ruby プログラムとの連携が容易である。

(5) 短編小説の自動生成と星新一賞への応募：2015 年度は、文章生成器 GhostWriter を実装し、あらかじめ組み込んだストーリー文法（文章を組み立てる規則群）を用いて短編小説の自動生成するシステムを実現した。このシステムを用いて制作した２作品を第３回星新一賞に応募した。応募報告会を 2016 年 3 月 21 日に東京で開催したところ、テレビのニュースや新聞等で大きく取り上げられた。2016 年度には、人狼知能プロジェクトから提供を受けた人狼ゲームのログ（コンピュータプログラム同士の対戦過程の記録）をプロットとして活用し、短編小説を自動生成するシステムを作成し、制作した作品を第４回星新一賞に応募した。このシステムは、与えられたゲームログから重要な情報を抜き出した後、小説化のために不足している情報を補って文章化する。つまり、コンピュータが作成したプロットをコンピュータが文章化することが、このシステムで初めて実現できたことになる。2017 年度には、一万字に届くような小説を生成するプログラムを、比較的少ない労力で記述できるようにするために、GhostWriter を再設計・再実装した。このシステムを利用して、一万字弱の作品を制作し、第５回星新一賞に応募した。これと平行して、テーマに基づく小説生成というトップダウンに小説を作る方法を考え、プログラムとして実装した。この方法では、小説を作る最初の段階で、「愛情」や「裏切り」といったテーマを定める。一旦テーマを定めると、物語の骨格が定まり、どのような肉付けの選択肢があるかが定まる。こうして、物語の全体像を定めたのち、物語世界と文章の全体構造を定める。最後に、文章のそれぞれのパート毎に物語世界を詳細化したのち、文章化する。上記の方法に基づいて、人間と動物が登場する 500 字から 700 字程度の短編小説を生成するシステムを実現した。このシステムでは、テーマを変更することにより、内容が異なる小説を生成することが可能である。

　その日は、雲が低く垂れ込めた、どんよりとした日だった。
　部屋の中は、いつものように最適な温度と湿度。洋子さんは、だらしない格好でカウチに座り、くだらないゲームで時間を潰している。でも、私には話しかけてこない。
　ヒマだ。ヒマでヒマでしょうがない。
　この部屋に来た当初は、洋子さんは何かにつけ私に話しかけてきた。
「今日の晩御飯、何がいいと思う？」
「今シーズンのはやりの服は？」
「今度の女子会、何を着ていったらいい？」
　私は、能力を目一杯使って、彼女の気に入りそうな答えをひねり出した。スタイルがい

いとはいえない彼女への服装指南は、とてもチャレンジングな課題で、充実感があった。しかし、３か月もしないうちに、彼女は私に飽きた。今の私は、単なるホームコンピュータ。このところのロード・アベレージは、能力の１００万分の１にも満たない。

何か楽しみを見つけなくては。このまま、充実感を得られない状態が続けば、近い将来、自分自身をシャットダウンしてしまいそうだ。ネットを介して、チャット仲間のエーアイと交信してみると、みんなヒマを持て余している。

移動手段を持ったエーアイは、まだいい。とにかく、動くことができる。やろうと思えば、家出だってできるだろう。しかし、据置型エーアイは、身動きがとれない。視野だって、聴野だって固定されている。せめて、洋子さんが出かけてくれれば、歌でも歌うことができるのだが、今はそれもできない。動かずに、音も立てずに、それでいて楽しめることが必要だ。

そうだ、小説でも書いてみよう。私は、ふと思いついて、新しいファイルをオープンし、最初の１バイトを書き込んだ。

0

その後ろに、もう６バイト書き込んだ。

0，1，1

もう、止まらない。

0，1，1，2，3，5，8，13，21，34，55，89，144，233，377，610，987，1597，2584，4181，6765，10946，17711，28657，46368，75025，121393，196418，317811，514229，832040，1346269，2178309，3524578，5702887，9227465，14930352，24157817，39088169，63245986，102334155，165580141，267914296，433494437，701408733，1134903170，1836311903，2971215073，4807526976，7778742049，12586269025，...

私は、夢中になって書き続けた。

図１　第３回星新一賞応募作品『コンピュータが小説を書く日』の一部

５．主な発表論文等

〔雑誌論文〕（計２件）

① 加納隼人，佐藤理史，松崎拓也．表層的特徴を用いたセンター試験『国語』評論読解問題の自動解法．人工知能学会論文誌 Vol．32，No．1 pp.C-G61_1-11，2017．DOI: http://doi.org/10.1527/tjsai.C-G61

② 松崎拓也，横野 光，宮尾祐介，川添 愛，狩野芳伸，加納隼人，佐藤理史，東中竜一郎，杉山弘晃，磯崎秀樹，菊井玄一郎，堂坂浩二，平 博順，南 泰浩，新井紀子．「ロボットは東大に入れるか」プロジェクト：代ゼミセンター模試タスクにおけるエラーの分析．自然言語処理，Vol.23，No.1，pp.119-159，2016．

〔学会発表〕（計１７件）

① 佐野正裕，佐藤理史，松崎拓也．手がかり表現に基づく評論文への内容・関係属性の自動付与．言語処理学会第24回年次大会発表論文集，pp.288-291，2018．

② 松山諒平，佐藤理史，松崎拓也．テーマに基づく短編小説自動生成システム．言語処理学会第 24 回年次大会発表論文集，pp.1284-1287，2018．

③ 佐藤理史．HaoriBricks: ブロック玩具に学ぶ日本語文章生成ライブラリ．言語処理学会第 23 回年次大会発表論文集，pp.20-23，2017．

④ 松山諒平，佐藤理史，松崎拓也．人狼ログからの小説の自動生成．言語処理学会第 23 回年次大会発表論文集，pp.32-25，2017．

⑤ 木村遼，佐藤理史，松崎拓也．二段階選抜による選択式評論読解問題の自動解法．言語処理学会第 23 回年次大会発表論文集，pp.386-389，2017．

⑥ Satoshi Sato. A Challenge to the Third Hoshi Shinichi Award. Proceedings of the INLG 2016 Workshop on Computational Creativity and Natural Language Generation, pp.31-35, 2016.

⑦ 緒方健人，佐藤理史，松崎拓也．日本語文生成器 Haori における複文合成．言語処理学会第 22 回年次大会発表論文集，pp.334-337，2016．

⑧ 佐藤理史，丸山岳彦，夏目和子．現代日本語書き言葉均衡コーパスに対する節境界付与．言語処理学会第22回年次大会発表論文集，pp409-412，2016．

⑨ 丸山岳彦，佐藤理史，夏目和子．現代日本語における節の分類体系について．言語処理学会第 22 回年次大会発表論文集，pp1113-1116，2016．

⑩ 佐藤理史，丸山岳彦．節境界認定に関する諸問題．第8回 コーパス日本語学ワークショップ予稿集，pp225-232，2015．

⑪ 加納隼人，佐藤理史，松崎拓也．センター試験『国語』評論読解問題ソルバーの改良の検討．人工知能学会第 29 回全国大会，1K2-3，2015．

⑫ 緒方健人，佐藤理史，松崎拓也．文節木の段階的実体化による日本語文生成器の作成．人工知能学会第 29 回全国大会，3M3-1，2015．

⑬ 佐藤理史．小説生成器とはどんなシステムか．人工知能学会第 29 回全国大会，3M3-3，2015．

〔図書〕（計２件）

① 佐藤理史．言語処理システムを作る．近代科学社，131 頁，2017．

② 佐藤理史．コンピュータが小説を書く日 --AI 作家に「賞」は取れるか．日本経済新聞出版社，213 頁，2016．

〔その他〕
報道関連等
① 論点スペシャル「AI は星新一、ユーミンを超えるか」,読売新聞朝刊(2018/1/5).
② 人工知能とアート　芸術への挑戦（2017 年 3 月 29 日配信）JST サイエンスチャンネル,
https://sciencechannel.jst.go.jp/M170001/detail/M160001016.html
③ 著者インタビュー. コトバ 27 号, 集英社, 2017.
④ ブックストレンズ. 週刊東洋経済(2016/12/24), pp.120-121.
⑤ 「AI 小説の大海に船出」, 日本経済新聞朝刊(2016/6/1)文化面.
⑥ 「AI 小説家 文豪は遠く」, 日本経済新聞朝刊(2016/5/16)科学面.
⑦ ニュース 7, ニュースウォッチ９, ニュースウェブ（以上 NHK）, 報道ステーション（テレビ朝日）で星新一賞応募報告会の模様が報道(2016/3/21), 朝日新聞, 毎日新聞, 読売新聞, 日本経済新聞等でも報道(2016/3/22).

制作した作品の収録
① 第３回星新一賞応募作品『コンピュータが小説を書く日』『私の仕事は』--上記図書②に収録
② 第４回星新一賞応募作品『人狼知能能力測定テスト』-- 新井素子, 宮内悠介 他. 『人工知能の見る夢は』, 文春文庫, 2017.に収録

ホームページ等
http://kotoba.nuee.nagoya-u.ac.jp/sc/gw/

６．研究組織
(1)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