JAIST Repository
https://dspace.jaist.ac.jp/

| | |
|---|---|
| Title | テキストの可読性を高める構造的書換えモデルの研究 |
| Author(s) | 島津，明 |
| Citation | 科学研究費助成事業研究成果報告書：1-5 |
| Issue Date | 2018-05-31 |
| Type | Research Paper |
| Text version | publisher |
| URL | http://hdl.handle.net/10119/15414 |
| Rights | |
| Description | 挑戦的萌芽研究，研究期間：2015～2017，課題番号：15K12094，研究者番号：60293388，研究分野：自然言語処理 |

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

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　３１　日現在

機関番号：　１３３０２
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１２０９４

研究課題名（和文）テキストの可読性を高める構造的書換えモデルの研究

研究課題名（英文）Study on a structural paraphrase model for improving the readability of texts

研究代表者
島津　明（Shimazu, Akira）

北陸先端科学技術大学院大学・その他・名誉教授

研究者番号：６０２９３３８８

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究は、テキストによる情報伝達の高度化に向けて、可読性を高める構造的書換え法を明らかにすることを目的とする。従来の研究が単語数や文数などに着目するのに対し、文の並び方や論理構造などのテキスト構造に着目する。国民年金法の条項などを分析し、要件効果構造を明確にし機械的に書換えるという視点から構造的書換えの枠組みを決めた。同法の約300条項を対象に、元文、構造的書換え、形態素構造、要件効果構造などからなるコーパスを作成した。4段階からなる構造的書換え法を提案し、前半部に機械学習を適用し、実験により可能性を示した。可読性の被験者実験を行い構造的書換えの有効性及び問題点を確認した。

研究成果の概要（英文）：This study aims to clarify a method for paraphrasing texts structurally to improve the readability towards advanced text-based communication. The study focuses on text structures such as sentence arrangement and logical structure, whereas past studies focus on the number of words, sentences and so on. Analyzing paragraphs of National Pension Act mainly, we formed a structural paraphrase framework from the viewpoint of clarifying requisite-effectuation structures and paraphrasing mechanically. We made a corpus containing original texts, structural paraphrases, morphological structures, requisite-effectuation structures and so on, covering about 300 paragraphs among the Act. We proposed a structural paraphrase method consisting of four stages, showed the possibility by the experiment applying machine learning to the first half stages. We also confirmed the effectiveness of the structural paraphrases by the subjective experiment and the problems in the experiment.

研究分野：　自然言語処理

キーワード：　書換え　テキスト構造　可読性　自然言語処理　法令工学

１．研究開始当初の背景

テキストの可読性（readability）の定義は研究者により異なるが、概ね、読み取れる性質、読み取れる度合いを表すといえる。legibility（layout, typeface）とは違うとする見方もあるが、理解しやすいテキストのあり方を追求する立場から、本研究は layout も考慮する。可読性に関する研究は英語が主で、語数、文長、態、埋込み、前置詞句などとの関連が調べられている。語数や音節数による単純な評価式も得られている。テキストの難易度が高いか低いか、内容が物語か説明か、トピックが日常か非日常かといった要素で調べる実験研究などもある。これらに対し、本研究は、従来、着目されていないテキスト構造に着目する。

本研究を着想したきっかけの１つは、科研B「電子化された情報の動的説明法の研究」(1998-2001) である。話し言葉には音声という一過性の制約があるが、その制約が談話構造と細かな発話単位の構造で補われている点に着目した。そこからテキスト構造が理解の容易さにつながるという発想を得た。もう１つのきっかけは、21 世紀 COE プログラム「検証進化可能電子社会」(2005～2009)（片山他、電子社会と法令工学、人工知能学会誌, 23, 4, 2008）である。このプログラムの中で、研究代表者等は、国民年金法の条項について、挿入文、長い名詞並列句への対処などに着目し人手で構造的書換えを行った（島津、国民年金法の構造的書換え、JAIST Press、2009)。その後、被験者実験を行い可読性が高くなることが分った（島津、法令工学：安心な社会システム設計のための方法論−法令文書の解析を中心に−, IEICE FR, 5, 4, 2012)。このような研究により、テキスト構造により１次元制約のくびきをいかに開放するか明らかにする本研究の着想を得た。

２．研究の目的

本研究は、１次元制約のあるテキストによる情報伝達の高度化に向けて、テキスト構造の違いによる可読性の違い、評価法、可読性を高める構造的書換えモデルを明らかにすることを目的とする。テキスト構造とは、テキストにおける文の並び方、談話標識、論理的関係、並列性などに関する構造である。１次元の文字列による制約をテキスト構造により克服することを目指す。例えば、文が並んだだけの構造に対し、段落や見出しを加えると可読性がよくなる。従来、語数や文長などと可読性との関係などに関する研究があったが、テキスト構造を対象にしておらず、可読性の捉え方が不十分など、問題がある。これに対し、本研究は、構造的書換えに着目し、評価法を明らかにして、機械学習に基づく構造的書換え法を明らかにする。

第１８条第２項第１文

| | |
|---|---|
| 要件部: | 年金給付は、その支給を停止すべき事由が生じたときは、 |
| 効果部: | その事由が生じた日の属する月の翌月からその事由が消滅した日の属する月までの分の支給を停止する。 |

図１　要件効果構造

３．研究の方法

可読性を高める構造的書換え法を明らかにすることを目的に，以下のように研究した。

・書換えるテキストは予算を考慮し主に国民年金法の条項を対象とした。一般性の視点から、新聞の経済記事などの構造的書換えについても検討した。

・条項の要件効果構造（図１）が明確になること、機械的に書換えができることを前提に、どのような構造的書換えが可能か、条項を分析した。

・分析に基づき、構造的書換えを定義し、予算の範囲で人手により注釈コーパスを作成した。

・構造的書換えの定義及びコーパスに基づき、機械的に構造的書換えを行う方法を明らかにし、特に、要件効果構造の解析法を具体化し実験し評価した。

・元テキストと構造的書換えの可読性をみる被験者実験を行った。被験者実験のために PC 上にツールを作成した。

・被験者実験の結果を定義やコーパスの修正に適用することは時間と予算の制約から将来の課題とした。

第１９条第２項

（元文）

前項の場合において、死亡した者が遺族基礎年金の受給権者であったときは、その者の死亡の当時当該遺族基礎年金の支給の要件となり、又はその額の加算の対象となっていた被保険者又は被保険者であった者の子は、同項に規定する子とみなす。

（構造的書換え）

前項の場合に、死亡した者が遺族基礎年金の受給権者であったときは、
Ａ 又は Ｂ は、前項に規定する子とみなす。

Ａ：　その者の死亡の当時、遺族基礎年金の支給の要件となる子

Ｂ：　その額の加算の対象となっていた 被保険者 又は 被保険者であった者 の子

図２　構造的書換え

第５２条の２第１項

（元文）

死亡一時金は、死亡日の前日において死亡日の属する月の前月までの第一号被保険者としての被保険者期間に係る保険料納付済期間の月数、保険料四分の一免除期間の月数の四分の三に相当する月数、保険料半額免除期間の月数の二分の一に相当する月数及び保険料四分の三免除期間の月数の四分の一に相当する月数を合算した月数が三十六月以上である者が死亡した場合において、その者に遺族があるときに、その遺族に支給する。ただし、老齢基礎年金又は障害基礎年金の支給を受けたことがある者が死亡したときは、この限りでない。

（構造的書換え）

死亡一時金は、死亡日の前日において、
死亡日の属する月の前月までの第１号被保険者としての被保険者期間に係る
Ａ ＋ Ｂ ＋ Ｃ ＋ Ｄ が３６月以上である者が死亡した場合に、
その者に遺族があるときに、遺族に支給する。
ただし、老齢基礎年金 又は 障害基礎年金 の支給を受けたことがある者が死亡したときは除く。

Ａ：　保険料納付済期間の月数

Ｂ：　保険料４分の１免除期間の月数の４分の３に相当する月数

Ｃ：　保険料半額免除期間の月数の２分の１に相当する月数

Ｄ：　保険料４分の３免除期間の月数の４分の１に相当する月数

図３　構造的書換え

| |
|---|
| ... 場合において(、) → ... 場合に(、) |
| ... 場合には、→ ... 場合、 |
| ... 場合は、→ ... 場合、 |

図４　類似の言い回しの置き換え

図５　被験者実験の PC 画面

４．研究成果

(1)国民年金法条項の構造的書換えの枠組み

国民年金法の条項を分析し構造的書換えの枠組みを得た。これは条項の要件効果構造を理解しやすくすること、及び機械化しやすくすることを考慮したものである(図２、図３)。枠組みは、かっこ書き挿入文の注への置き換え、要件効果構造の主題や要件を考慮した改行、複雑な名詞句の英字記号による置き換え、並列名詞句の要素ごとへの余白文字の挿入、算術式への置き換え、類似の言い回しの標準的なものへの置き換えなどからなる。構造的書換えの枠組みは、本研究の代表者らが過去に行った国民年金法の書換えを手がかりに、過去に検討した他の構造的書換えも考慮し、研究協力者とともに、構造的書換えの様々な

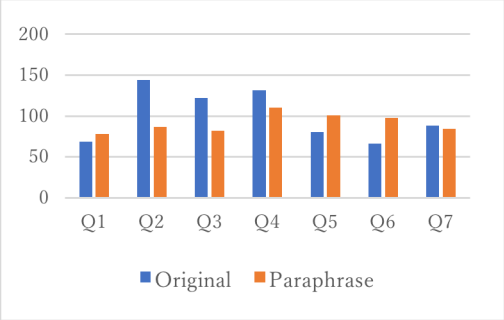


図６　被験者実験の結果

可能性を検討しながら、国民年金法の条項を再分析して得たものである。過去に行った国民年金法の書換えは、見やすくしようと直感的に試行錯誤して行ったもので、機械化を前提にしたものではなかった。過去に検討した他の構造的書換えというのは、論理記号や線図形などの利用である。これの被験者実験は理解が容易になるか判然としない結果であった。時間や予算を考慮し、今回の構造的書換えでは、論理記号などは試みなかった。

(2)国民年金法条項の構造的書換えコーパス

構造的書換えの枠組みに基づいて、国民年金法の主要部分である第１章から第９章について、約 300 条項の構造的書換えコーパスを作成した。各条項について、元文テキスト、構造的書換え、それぞれの形態素構造、主題部・要件部・効果部の注釈などの内容がある。

(3)構造的書換え法

構造的書換えの枠組みに基づき、4 段階からなる方法を提案した。処理の要点は、条項の要件効果構造の明確化に基づく構造的書換えへの変換ある。具体的には、４段階は、要件効果構造の認識、節への分割、標準的な言い回しへの置き換え、構成素から構造的書換えへの変換である。なお、書換え処理の前に、括弧挿入文は本文から外し注としておく。４段階のうち第１段と第２段については、深層学習に基づく方法を提案した。第１段階の要件効果構造の認識は、BI-LSTM-CRF という RNN（Recurrent Neural Networks）に基づくモデルにより行う。認識を系列ラベリングタスクとして扱い、文節列を入力として、要件部と効果部を出力する。第２段階の節への分割は、第１段と同様の系列ラベリングタスクとして、文節列を入力に分割点を出力とする。第１段と第２段については、構造的書換えコーパスを利用して実験を行い、第１段は約 80%、第２段は約 85%程度の精度を実現し、見込みがあることを確認した。第３段、第４段については、類似の言い回しを標準的なものに置き換える規則をまとめるとともにアルゴリズム化を検討した。類似の言い回しを置き換える規則は 115 ある（図４）。

(4)被験者実験

被験者がテキストを読んで質問に回答する時間を計測し、テキストの可読性を評価する実験を試みた。被験者は PC の画面に表示される質問を見て、紙に印刷された元文または構造的書換えの条項を読み、条項に関する質問の回答を PC に入力した（図５）。被験者は、正解と考えるテキストを回答として入力した。質問が画面に表示されてから回答が終わるまでの時間を計測した。被験者は元文、構造的書換え、それぞれ 12 名であった。8 名が大学や大学院の学生、4 名が卒業生で、平均年齢は 24 歳であった。質問は 7 つであった。実験に用いた構造的書換えはコーパスから選んだものである。

統計的検定により、時間差の有意差がない場合が 4 つ（図６の P1、P4、P5、P7）、構造的書換えが速い場合が 2 つであった(図６の P2、P3)。不正解は、元文が 15、書換えが 13 あった。構造的書換えが多少よい。構造的書換えが有意に速いのは、構造的書換えが算術式を利用した場合であった。元文が有意に速かったのは、元文が短い文で、構造的書換えの行数が 2.5 倍の場合であった。可読性に有意な差がない場合は、質問内容と回答の仕方が影響したとみられる。質問は、対象条項の意味

表１　被験者に対する質問

| |
|---|
| （問２）<br>第５条第２項は「保険料納付済期間」を定義しています。保険料納付済期間は、３つの期間を合算したものです。一つは、第７条第１項第１号に規定する被保険者としての被保険者期間のうち納付された保険料に係る被保険者期間です。もう一つは第７条第１項第３号に規定する被保険者としての被保険者期間です。後一つは何でしょうか？ |
| （問 7）<br>第４９条第１項は、寡婦年金について書いています。夫が死亡した場合に、妻に寡婦年金が支給されますが、そのためには、夫及び妻にそれぞれ条件が必要です。夫の条件は概略、以下です。<br>・ 第１号被保険者としての被保険者期間に係る保険料納付済期間と保険料免除期間とを合算した期間が二十五年以上。<br>・ 保険料納付済期間 又は 納付することを要しないものとされた保険料に係る期間以外の保険料免除期間を有する者。<br>・ 夫が障害基礎年金の受給権者であったことがない。<br>・ 夫が老齢基礎年金の支給を受けていない。<br>妻の条件の一部は、以下です。<br>・ 夫の死亡当時、夫によって生計を維持していた。<br>・ 夫との婚姻関係が十年以上継続していた。婚姻関係は、婚姻届出をしていないが、事実上婚姻関係と同様の事情にある場合も含む。<br>妻の残りの条件は何でしょうか？ |

内容の理解というより、要件部の要素である要件や算術式の項に対応する節や句がどれかを問うものが主であった（表１）。要件部の要素の箇所がどれであるかという問いは被験者が慣れていないこととみられ、回答の把握とテキスト入力に手間取ったようであった。この推察は、過去に行った被験者実験（A. Shimazu, Structural Para-phrase of Law Paragraphs, Jurisin2017）との対比によるものである。過去に行った被験者実験は、条項の意味内容について問い、回答は選択肢の中から選ぶ方式で、8 問中 7 問について構造的書換えが有意に速かった。これらのことから質問内容や回答方式が結果に影響していると考えられる。

５．主な発表論文等

〔雑誌論文〕（計２件）


(1) Son Truong Nguyen, Minh Le Nguyen, Satoshi Tojo, Ken Satoh, Akira Shimazu. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. Artificial Intelligence and Law, 査読有、DOI: 10.1007/s10506-018-9225-1, 2017, pp.1-31.

(2) Tho Thi Ngoc Le , Minh Le Nguyen, Akira Shimazu. Unsupervised Keyphrase Extraction: Introducing New Kinds of Words to Keyphrases. AI 2016: Advances in Artificial Intelligence, 査読有、Volume 9992 of the series Lecture Notes in Computer Science, pp.665-671, 2016.


〔学会発表〕（計４件）


(1) Son Truong Nguyen, Minh Le Nguyen, Akira Shimazu and Kiyoaki Shirai. Structural Paraphrasing in Japanese Legal Texts. Eleventh International Workshop on Juris-informatics (JURISIN 2017), 筑波大学東京キャンパス（東京都）、2017.11.14.

(2) Akira Shimazu. Structural Paraphrase of Law Paragraphs. Eleventh International Workshop on Juris-informatics (JURISIN 2017), 筑波大学東京キャンパス（東京都）、2017.11.14.

(3) Son Trong Nguyen, Minh Le Nguyen, Ken Satoh, Tojo Satoshi, Akira Shimazu. Single and multiple layer BI-LSTM-CRF for recognizing requisite and effectuation parts in legal texts. The 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts, London (UK), 2017.6.16.

(4) Son Trong Nguyen, Minh Le Nguyen, Ho Bao Quoc, Akira Shimazu. Recognizing logical parts in legal texts using neural architectures. The Eighth International Conference on Knowledge and Systems Engineering (KSE 2016), Hanoi (Vietnam), 2016.10.7


〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕

６．研究組織

(1)研究代表者
島津　明（SHIMAZU AKIRA）
北陸先端科学技術大学院大学・その他・名誉教授
研究者番号：６０２９３３８８

(2)研究分担者
グエン レ ミン（Nguyen Minh Le）
北陸先端科学技術大学院大学・先端科学技術研究科・准教授
研究者番号：３０５０９４０１

(3)連携研究者

(4)研究協力者