科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　３０　日現在

機関番号：　１２６０８
研究種目：　若手研究(A)
研究期間：　2011～2013
課題番号：　２３６８００１４

研究課題名（和文）談話解析技術に基づいた文章推敲支援

研究課題名（英文）Writing support for revising texts based on discourse analysis

研究代表者
飯田　龍（Iida, Ryu）

東京工業大学・情報理工学（系）研究科・助教

研究者番号：４０４６４２７６

交付決定額（研究期間全体）：（直接経費）　　7,300,000円、（間接経費）　　2,190,000円

研究成果の概要（和文）：本研究では，説得力のある説明的な文章の作成するために文章の推敲を支援する技術を開発した．まず，談話的な観点から推敲を行う手順を提案し，その手順にしたがって日本語の教師と日本語小論文添削の専門家を6名が日本語母語話者が作成した120編の小論文を推敲することで正解データを作成した．次に，推敲支援の問題を(1)文章のつながりの良さの推定，(2)談話単位の並び換え，(3)局所的な表現の修正の3つの問題に分解して，各問題を解くことで推敲支援を実現する．

研究成果の概要（英文）：In this research project, we developed a technique of supporting to revise texts for writing persuasive and informative texts. We first proposed a procedure for revising a text from a discourse perspective. Following the procedure six revisers who have enough experience in either teaching Japanese or scoring Japanese essays revised 120 Japanese essays written by Japanese native speakers. We then decomposed the text revision task into the following three subtasks: (1) evaluating text coherence, (2) reordering discourse units in a text and (3) modifying expressions for improving local coherence. We achieved supporting text revision by automatically solving each subtask.

研究分野：　総合領域

科研費の分科・細目：　情報学・知能情報学

キーワード：　推敲支援　自然言語処理　談話解析　首尾一貫性

１．研究開始当初の背景

計算機を用いた文章作成・推敲支援など，教育的応用を指向した研究が盛んに行われている．近年，これらの研究課題は機械学習アルゴリズムの発展により，語や語の並びなどの表層的な手がかりを用いて小論文の自動評価や助詞など単語レベルの誤用検出などの研究が進められている．これらの技術の進歩に対し，大学(院)生の卒業論文・修士論文などを実際に推敲する立場で記述された内容を分析してみると，書くべき内容は充足しているものの，文と文のつながりや，文内の構成素の配置が適切ではないため，結果的に可読性の低い文章となっていることが多い．これは日本では中学・高校のカリキュラムに文章作成の授業が無いこと，また特に理工学系の学生は一般的に大学入試に小論文が含まれないために，これまでに文章作成に関する指導を一切受けていないことに起因する．また，留学生が作成した日本語の文章を見ると，主語が過剰に省略されるなど，参照表現生成における問題も存在する．しかし，既存の文章作成・推敲支援のシステムでは前述の通り表層的な手がかりを主に利用しているため，これらの問題を改善することはできない．英語を対象に同様の問題を見た場合には，特に Educational Testing Service の研究者が英語学習支援における談話構造の情報の利用について議論しているが，彼らの研究の方向性は自動採点であり，どのように文章を作成すべきかを支援する立場とは異なる．このように研究開始当初は，計算機による推敲の支援，特に談話レベルの推敲に関しては研究として未着手の状態であった．

２．研究の目的

本研究課題では，文章のつながりの良さの推定，より適切な文の順序，代名詞などの参照表現の修正・省略などの自然言語処理の問題を解くことで，説得力のある説明的な文章の作成を支援することを目的とする．

本研究では文章を推敲の手続きを，(a)適切な語彙の選択など内容レベルの推敲と(b)結束性レベルの談話のつながりの良さの 2 つに分解して考える．前者に関する自動推敲支援を考えた場合，推敲システムは記述する内容自体を把握して支援する必要があるため非常に困難な課題設定となるが，これに対し，後者の問題では，主題を表す表現が文章中でどのように連鎖するのか，どの箇所にどのような接続表現を挿入するする必要があるかなど，文章の内容とは独立した談話レベルの手がかりを利用することで一般的に利用可能な推敲支援が実現できる可能性がある．そこで，本研究課題では特に(b)の談話レベルの推敲支援技術の実現に着目し，日本語を対象にした推敲支援技術の確立を目指す．

３．研究の方法

本研究で実現する推敲支援のために，まず文章のつながりの良さの改善の余地のある文章集合を収集し，それを日本語小論文の添削者などの専門家に修正させることで，推敲の正解データを作成する．この際，作業者が修正を行う場合に，何を基準にどの箇所をどのように編集したかについても情報を付与することで，どのような手がかりをもとに自動修正を行えばよいかという点について調査を容易にする．この調査によって明らかになった事実に基づいて，文の順序並び換えや参照表現生成の課題を解く．

一方で，推敲の正解データの作成は修正のガイドラインの決定に時間がかかるため，本研究で対象とする照応・省略関係がすでにアノテーションされたコーパスも併用することで，研究を効果的に進める．具体的には，省略関係と名詞句共参照関係がすでにアノテーション済みである NAIST テキストコーパスを利用することで，参照表現生成や文の並び換え問題を扱う．

４．研究成果

本研究では文章推敲支援の課題をその正解データの作成と，(1)文の並び換え，(2)参照表現の修正・省略，(3)文章のつながりの良さの推定の 3 つの問題を解くことで実現する．以降で，各タスクへの取り組みの内容とその結果について報告する．

(1)**推敲の正解データの作成**：推敲の正解データの作成を考えた場合，まずどのような文書を対象に正解を作成することが本研究で対象とする談話レベルの推敲を評価するのに適しているかを考える必要がある．例えば，新聞記事などの専門家が記述した文書データを利用した場合，そこの修正すべき内容が含まれないことになり，一方，日本語学習者が作成した文書データの利用を考えた場合，文法誤りやスペルミスなど，他の種類の誤りのために，より高次の談話レベルで文章を改善することを考えることが困難となる．このため，文法誤りなどの局所的な誤りが比較的少なく，かつ，談話レベルでは改善の余地がある文書を収集することを考える必要がある．このため，日本語を母語とする人材で，かつ，中学生や高校生のような文章作成に不慣れな人材が記述した文書を対象とすることが望ましいと考えた．そこで，本研究では，この条件に合致する文書をすでに収集している研究者の協力を仰ぐことで，推敲の対象となるデータを収集した．具体的には，宇佐美(2009)が収集した，高校生が「日本における英語の早期教育の是非」について書いた小論文 120 編を対象に作業者が修正を行う．事前に行なった修正作業の結果，単純に修正を依頼するだけでは，作業者は局所的な文法誤り・スペルミスの修正しか行わない傾向にあることがわかったため，下記に示す作業手順にしたがって，作業者は修正作業を行った．

- 節などの談話単位への分割
- 談話単位の文章の論理的な構成素（言説構成素）へのまとめあげと並び換え
- 参照表現・接続表現の局所的な修正

修正の具体例を以下に示す．この例のように，作業者はある言説構成素に the（著者の主な主張）や main（主張の理由）などの論理的な構成に関するラベリングを行うことで，その構成素をどこに配置するかを吟味し，結果的に文の位置を修正していることがわかる．

	ラベル	談話単位ID	修正結果
0	the	(7)	〈ただし｜〉小学校における英語の早期教育〈が｜は〉必要である〈という｜〉。
1	main1	(5)	〈そのような意味では、｜〉中学校や高校で英語を学習して、「英語が難しい」と苦手意識を持ってしまう前に、小学校で「英語が楽しい」と思えるような教育をするの〈が｜は〉、むしろ必要なことと〈｜と思う〉。
		(6)	日本語と同様に、相手とのコミュニケーションをとる手段として早期から英語に触れていれば、後になってから苦労して学ぶということに〈｜は〉ならない。
2	back1	(0)	私〈が｜は〉小学校中学年、高学年のときに英語に触れる機会があった。
3	elab1	(1)	それ〈が｜は〉中学校や高校における「英語の授業」という〈。｜よりは、〉
		(2)	遊びの感覚で楽しめる〈。｜〉
		(3)	ものであった。
4	supp1	(4)	例えば、英語の歌を歌ったり、朝の健康観察のときに先生と英語であいさつをしたり、英語を身近に感じることができるものであった〈｜と思う〉。
5	main2	(8)	〈前術の通り｜ただし、〉子どもたちが「楽しい」と思えるよう〈だ。｜な〉
		(9)	ものでない〈。｜と〉
		(10)	〈｜英語の早期教育は〉全く意味がなく、むしろ逆効果になってしまう。
6	supp2	(11)	私〈が｜は〉幸運なことに、小学校だけでなく中学校で〈｜も〉英語の授業が「楽しい」ものであると思うことができた。
		(12)	それ〈が｜は〉何よりまず、当時の ALT の先生のおかげであったかもしれない。
		(13)	というのも、彼女と私に〈｜は〉共通の趣味があり、よくそのことについて話したり、英語の授業で洋楽を聴いたり、英語で書かれた レシピを見ながらクッキーを作ったりと、私だけでなく誰が楽しめるような時間にしてもらえた〈｜からである〉。
7	reb	(18)	「そんなに早くから子どもに英語を学ばせる必要〈が｜は〉ない」という人〈が｜も〉いるであろう〈。｜が、〉
8	solu	(19)	その早い時期に子どもが楽しんで英語に触れることができれば、それが子どもの可能性を広げるということに繋がるのではないだろうか。
9	conc	(14)	英語に限ったことではない〈。｜が、〉
		(15)	何かを学ぶということにおいて、一番大切なの〈が｜は〉、学ぶ本人がどれぐらい意欲を持〈つ。｜って〉
		(16)	学べるかということである。
		(17)	「もっと知りたい」という意欲が何より本人の力を伸ばすのである。

小論文 120 編を 60 編ずつに分割し，各 60 編を対象に日本語の教師や添削の専門家がそれぞれ 3 名ずつ修正を行うことで約 354 編の推敲正解データを獲得した（360 編中 6 編は作業が中断された）．

この作業者が修正したデータを対象にどのような文の並び換えが起こっているのかを調査したところ，頻繁に起こる並び換えには，著者の主張を文章頭に配置する，主張の理由を述べた後にはその詳細や具体例を配置する，反論やその解決策は相対的に文章の後半に配置するという傾向が観察できた．この傾向を実際の自動並び換えに適用することで，よりつながりの良い文章へ改善できると考えられる．この研究内容を言語処理学会第 19 回年次大会で発表し，優秀賞を受賞した．

(2)**文の並び換え**：(1)で作成した 364 編の修

正データを対象に修正前から修正後の文の順序へ自動的に並び換える課題に取り組んだ．まず，この修正された小論文の特徴を調査するために，言説構成素のラベルや作業者間の作業の一致率を調査した．この結果，一つの小論文に割り当てられた 3 人の作業者はそれぞれ独立に筆者の意図を異なる粒度で読み取って文の順序を並び換えているため，順位相関のレベルでは一致率が低いことがわかった．ただし，この並び換えは作業者がアノテーションした言説構成素のラベルに依存して発生しているため，言説構成素のラベルを与えた上で並び換えを行う問題を考える．このラベルがすでに与えられた状態で書き換えを行うという手続きは小論文記述のオーサリングツールで論理的にどのような内容を記述するかを明示的に指定しながら文章を書く作業に相当するため，そのようなオーサリングツールの利用時に推敲の支援が可能になると考えられる．

ここで対象とする文の並び換えの問題は順位相関係数で 0.75 という高い値を持つ元文章から部分的に必要な箇所のみを適切な位置に再配置を行うという難易度が高い問題であり，元文章でどの位置に出現したかという情報が重要となる．このため，元文章の出現位置の情報を特徴量とし，さらに並び換えの対象となる言説構成素のラベルやその前後の情報を特徴量として利用することで文の並び換えを行う．また，元文章から修正後の文章へのリランキングの問題として扱うことで各特徴量の有効度はランキング学習を利用することで自動的に推定する．

人手修正データを用い，ここで提案したランキング学習に基づく手法の有効性を調査した．評価実験の結果，ランキング学習に基づく手法を採用して並び換えを行うことで修正前の順序と比較して，ケンドールの順位相関係数が向上することを示した．

(3) **参照表現の修正・省略**：文の並び換えを行った結果，代名詞や指示連体詞をともなう名詞句が冗長に記述される可能性があり，また一方で，文章のつながりとしては過剰に省略が起こるという場合も考えられる．このため，適切に参照表現を再生成する必要がある．そこで，文脈に応じて適切な参照表現を生成する技術を開発した．共参照連鎖（文章中で同じ実体を指す表現の連鎖）の各要素がどのような意味カテゴリに属するか，どのような文脈で出現しているかなどを特徴量とし，「は」をともなって主題化して生成する，主語/目的語として生成する，省略する，のいずれかを選択する 3 値分類問題として定式化した．この提案手法を評価するために，省略・共参照関係がアノテーションされた NAIST テキストコーパスを利用した評価を行った．この結果，規則ベースの生成手法の F 値が 0.291 であったのに対し，提案手法は 0.550 という高い F 値で生成できることを示した．また，日本語を母語とする 3 名の研究者に人手で参照表現を生成する課題に取り組んでもらい，3 名が同じ判断を行った事例のみを対象に提案手法を適用したところ，0.585 というさらに高い F 値を得た．生成を誤った事例を分析し，何について書かれているかを読み手が理解しやすい情報が省略されやすい（例：「日本」は文脈に比較的独立に省略されやすい）や話題の転換が再度名詞句として生成することに影響していることを明らかにした．このような文脈の情報を機械学習の特徴量として導入することは困難であるため，今後どのようにこの問題を解決すべきかを検討する必要がある．

(4) **文章のつながりの良さの推定**：(2)の文の並び換えの結果，元記事と比較して文章のつながりの良さが改善したことを保証する必要がある．このため，任意の文章のつながりの良さのスコアを計算する手法を開発する必要がある．この技術開発のため，本研究では「つながりの良い文章が書かれている場

合には適切に照応関係が多用される」という仮説にしたがい，照応関係を自動的に解析する技術を利用して，文章のつながりの良さのスコアを定義した．ここで開発したスコアの良さを評価するために，NAIST テキストコーパスを利用した評価を行った．既存研究と同様に，もとの文章中の文をランダムに並び換えた記事と元記事を入力として与え，後者がより高いスコアを出力した割合でそのスコアの良さを評価した．この結果，既存研究と比較して，提案するスコアを利用することでより適切につながりの良さを推定できることを明らかにした．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

1．　飯田龍．意味・談話処理課題の規格化とその諸問題．人工知能学会誌第27巻第3号 特集「ポスト経験主義の言語処理」, pp.318-325, 2012．

〔学会発表〕（計　14　件）

1．　飯田龍．アノテーションされた結果の人手分析 -ポストアノテーションと事例クラスタリング-．言語処理学会第 20 回年次大会ワークショップ「自然言語処理の発展に向けた情報共有・討論」．2014 年 3 月 21 日．北海道大学．

2．　飯田龍，徳永健伸．小論文推敲のための談話単位の並び換え．言語処理学会第 20 回年次大会，pp.89-92．2014 年 3 月 18 日.北海道大学．

3．　飯田龍，光田航，徳永健伸．アノテーション時の作業者の振舞いの収集とその分析．情報処理学会自然言語処理研究会予稿集，NL-213-1，pp.1-9，2013 年 3 月 18 日.北海道大学．

4．　宮原聡，飯田龍，徳永健伸．日本語書き言葉を対象とした談話単位分割基準の提案と自動分割の評価．情報処理学会自然言語処理研究会予稿集，NL-211-02，pp.1-7，2013 年 5 月 23 日．北陸先端科学技術大学院大学東京サテライト．

5．　*Ryu Iida, Takenobu Tokunaga. Automatic voice selection in Japanese based on various linguistic information. In Proceedings of the 14th European Workshop on Natural Language Generation (ENLG 2013), pp.147–151,* August 9, 2013. Sophia, Bulgaria.

6．　飯田龍，徳永健伸．談話的な手がかりを利用した日本語の節の受動化．言語処理学会第 19 回年次大会発表論文集，pp.662-665，2013 年 3 月 14 日．名古屋大学．

7．　飯田龍，徳永健伸．談話レベルの推敲支援のための人手修正基準．言語処理学会 第 19 回年次大会発表論文集，pp.830-833，2013 年 3 月 14 日．名古屋大学．

8．　Ryu Iida, Koh Mitsuda, Takenobu Tokunaga. Investigation of annotator's behaviour using eye-tracking data. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp.214–222, August 9, 2013. Sophia, Bulgaria.

9．　Ryu Iida, Takenobu Tokunaga. A metric for evaluating discourse coherence based on coreference resolution. In Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), pp.483–494, December 14,2012. Mumbai, India.

10．　飯田龍，徳永健伸．照応・共参照解析に基づく文章の首尾一貫性の指標．言語処理学会第 18 回年次大会発表論文集，pp.22-25，2012 年 3 月 14 日．広島市立大学．

11. 飯田龍，笹野遼平．日本語ゼロ照応関係に対する特徴分類とそのアノテーション．テキストアノテーションワークショップ・コンテスト，C05，pp.1-7，2012 年 8 月 6 日．国立情報学研究所．
12. 飯田龍，徳永健伸．アノテーション作業者の内省を顕在化するためのデータ収集．テキストアノテーションワークショップ・コンテスト，C04，pp.1-7，2012 年 8 月 6 日．国立情報学研究所．
13. 飯田龍，徳永健伸．日本語書き言葉を対象とした参照表現の自動省略 -人間と機械処理の省略傾向の比較-．情報処理学会自然言語処理研究会予稿集，NL-206-15，pp．1-8，2012 年 5 月 11 日．東京工業大学．
14. 飯田龍，徳永健伸．照応・共参照解析を利用した文章の首尾一貫性の評価．情報処理学会自然言語処理研究会予稿集，NL-204-11，pp.1-8，2011 年 11 月 21 日．石垣市商工会館．
15. Ryu Iida, Massimo Poesio. A Cross-Lingual ILP Solution to Zero Anaphora Resolution. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), pp. 804–813, June 21, 2011. Portland, USA.

〔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〕ホームページ等

http://www.cl.cs.titech.ac.jp/~ryu-i/

６．研究組織
(1)研究代表者
飯田 龍（Iida，Ryu）
東京工業大学・大学院情報理工学研究科・助教
研究者番号：40464276

(2)研究分担者
（　　）
研究者番号：

(3)連携研究者
（　　）
研究者番号：