言語処理技術を利用したテキストの一貫性向上のための推敲支援

Revision support for improving text coherence by using language processing technology

徳永 健伸(Tokunaga, takenobu, Ph. D.)
東京工業大学 情報理工学院 教授
(Professor, School of Computing, Tokyo Institute of Technology, PhD)
Association for Computational Linguistics, Association for the Advancement of Artificial Intelligence, Association for Computing Machinery, Cognitive Science Society, International Cognitive Linguistics Association, 人工知能学会, 言語処理学会, 計量国語学会 他
受賞:Outstanding Paper award at the 28th International Conference on Computational Linguistics (2020), Special Session Best Paper Award at the 9th International Conference on Knowledge and Systems Engineering (2017), Best Paper Award at TextGraphs-11: Graph-based Methods for Natural Language Processing (2017) 他
著書: 東工大英単–科学・技術例文集 新装版, 研究社 (2021), コーパスと言語処理, 朝倉出版 (2017), Handbook of Linguistic Annotation, Springer (2017), LMF-Lexical Markup Framework, John Wiley & Sons (2013) 他
研究専門分野:計算言語学 自然言語処理

あらまし

本研究では言語処理技術を利用して、よりよいテキストを書くための支援をすることを目的としている。「よい」テキストの条件には多くの要因がある。表記や文法の間違いを指摘する機能はすでに商用システムで実現されているが、テキストの構成や文間のつながりの自然さなどの改善を支援の対象としているものはない。我々は特にテキスト中の文の並びに注目し、まず、テキストの構造を解析し、その構造を元に文を適切に並び換えることによってテキストを改善するシステムの構築を目指している。そのための研究項目として、(1) テキストの構造とテキストを改善するための修正を記録した基礎データの構築、(2) テキストの構造の解析手法、(3) 文の並び換えによるテキストの改善手法を設定し研究をおこなった。結論として、テキスト構造解析の精度をさらに高めることによって、元テキストの文順序を自動的に並び換えてテキストを改善できる可能性があることを示した。

１．研究の目的

本研究は言語処理技術を用いてテキストの構造を解析し、テキストの一貫性を向上させる修正を助言することによって、筆者の推敲を支援する手法を実現することを目的としている。また、そのための基礎データとして、テキスト中の文と文の関係と同時に、テキストの一貫性を向上させるための構造レベルの修正をアノテーションしたコーパスを構築する。

２．研究の背景

テキストは人間の言語によるコミュケーションのための主要な手段である。そのためには論理的でわかりやすいテキストを書くことが重要となる。大学入試でも記述式問題の導入が検討されており、「よい」テキストを書く能力は教育的な観点からも注目を集めている。「よい」テキストの条件には多くの要因がある。スペルミスや簡単な文法のチェックなどの表層的な修正を支援する機能は Microsoft Word、Grammarly[1]、Ginger[2]などの商用のシステムにおいてすでに実現されている。また、研究レベルでは、外国語学習者の書いたテキストの誤り箇所に情報を付与したデータ(コーパス)から機械学習の手法を用いて誤りを自動訂正する研究がおこなわれている[3]。しかしながら、訂正の対象はいずれも一文内の局所的で表層的な誤りにとどまり、テキストの構成や連続する文間のつながりの自然さなどを修正の対象として扱っている研究はほとんどない。これに対して本研究では、テキストの一貫性を対象とする。テキストの一貫性とはテキスト中の文と文が意味的に関連するように配置されていることをいう[4]。一貫性は「よい」テキストであるためのひとつの要件であり、一貫性のあるテキストは、話題に意味的に関連する文が自然に連なるために読み易いものとなる。従来、テキストの読み易さはテキスト中の語の長さ、語の難易度、文長、構文などの表層的な性質によって計測されてきたが、我々は内容を考慮した一貫性に着目し、最終的にはこれを向上させる

言語処理技術を利用したテキストの一貫性向上のための推敲支援

Revision support for improving text coherence by using language processing technology

ための適切な助言を与えることを目指している。つまり、従来の研究が間違いを訂正することであったのに対し、我々はよりよいテキストに改善するための支援を目的とする。

３．研究の方法

本研究の目的を達成するために以下の 3 つの研究課題を設定し、研究を進めた。

(1) テキスト構造とテキスト構造レベルの修正情報を付与したコーパスの構築
(2) テキスト構造の解析手法
(3) テキスト構造を考慮した文順序の整列手法

3.1 テキスト構造とテキスト構造レベルの修正情報を付与したコーパスの構築

深層学習技術の発展により、多くの人工知能関連の研究分野では対象とする課題用の学習データを用意し、機械学習の手法を援用して課題を解決する手法が主流となっている。本研究でも機械学習の技術を利用することを前提とし、そのための基礎データとして、テキスト構造とそれをより一貫性の高いテキストに修正する操作をテキストに付与(アノーテション)したコーパスを構築した。テキスト構造を付与したコーパスを作成した研究はこれまでにもいくつか報告されているが、テキストの一貫性を向上させるための修正操作を合わせて付与したコーパスは皆無であった。

テキスト構造と修正操作を付与する元コーパスとしては、神戸大学で構築された ICNALE[5]を使用した。ICNALE には外国語として英語を学ぶアジア圏の大学生が書いた 400 語程度の英語エッセイが収録されている。各エッセイは言語的、内容構成的ないくつかの観点から評点が付与されている。また、文法的な誤りを修正したサブセットも用意されている。我々の感心がテキストの表層的な誤りや改善ではなく、大局的なテキスト構造の改善であることから、ICNALE の文法誤りを修正したサブセットから中程度の評点が付与された 434 エッセイをアノテーションの対象として選んだ。

これらのエッセイについて、まず、エッセイ中の文と文の間に(1)支持 (support)、(2)攻撃 (attack)、(3)詳細化 (detail)、(4)再言明 (restatement)の 4 つの関係を付与することによってテキスト構造をアノテーションする。図 1 に「公共の場での喫煙は法律で禁止すべきであるか」という問に対して、自分の意見を述べた

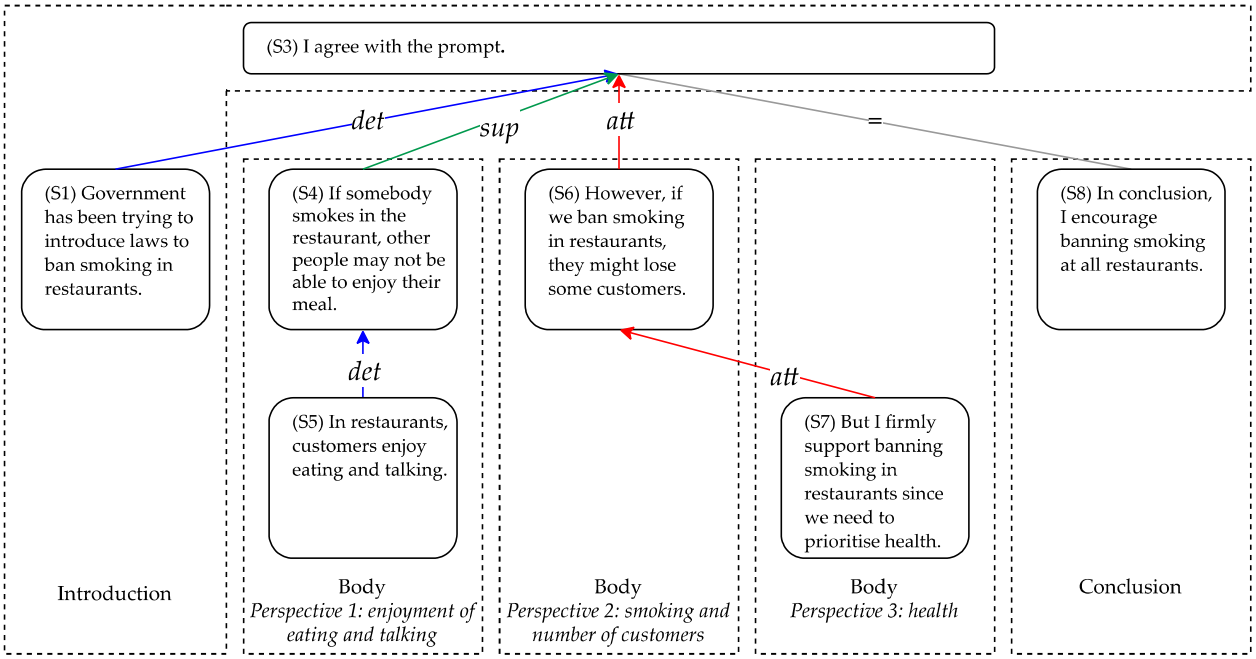

図 1　テキスト構造のアノテーション例

言語処理技術を利用したテキストの一貫性向上のための推敲支援

Revision support for improving text coherence by using language processing technology

エッセイにテキスト構造のアノテーションをした例を示す。文ラベル(Si)に付与された数字 i はエッセイ中の文の順序を示し、文と文の間には 4 つの関係のいずれかが付与されている。

さらに、これらのエッセイの一貫性を改善するために人手で修正した記録をアノテーションする。自由な修正を許すとデータ化が困難になるので、修正操作は文の順序の変更とそれにともなう必要最小限の接続詞や指示詞の変更に制限する。このようなアノテーションをおこなうためのガイドラインを策定した。このアノテーションガイドラインにしたがい、 20 エッセイについて 3 名のアノテータが独立にアノテーションし、その結果について一致度の分析をおこなった。この分析結果からアノテーションガイドラインを改訂し、これにしたがって、残りの 414 エッセイについて応用言語学の専門家 1 名によるアノテーションをおこなった。

我々のアノテーションは、テキスト中の単語に品詞を付与したり、語と語の係り受け関係を付与するようなアンテーションと比較して、各文の意味内容に加え、テキストを大局的に理解する必要があり、アノテーション作業がテキスト広範にわたることから、アノテーションの実施に先立って専用のアノテーションツール TIARA を開発した。TIARA は JavaScript で記述されており、Web ブラウザ上で動作するため特別なインストール作業は不要である。また、文間の関係はユーザ定義が可能なので、一般のテキスト構造のアノテーションにも使える汎用的なアノテーションツールとなっている。図 2、3 は TIARA のインタフェースである。

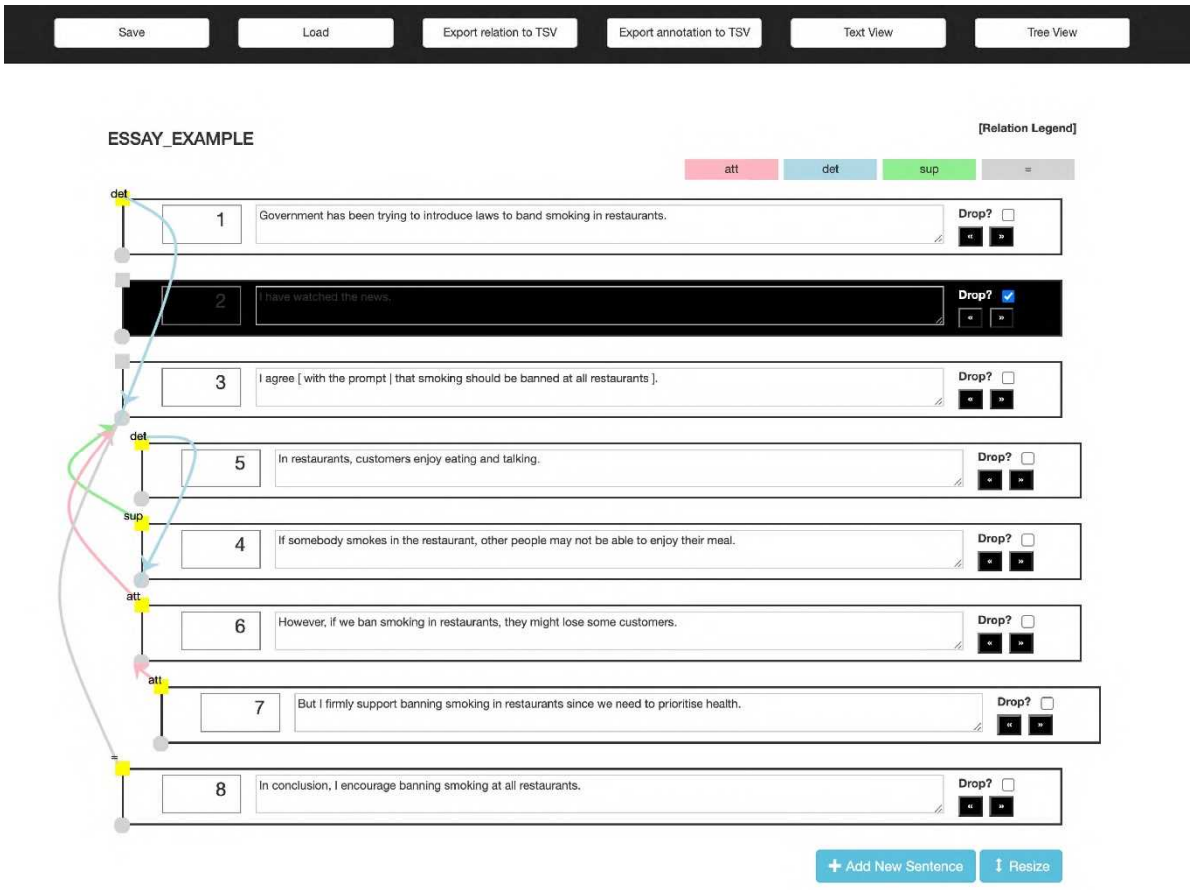

図 2　TIARA のテキストビュー

言語処理技術を利用したテキストの一貫性向上のための推敲支援

Revision support for improving text coherence by using language processing technology

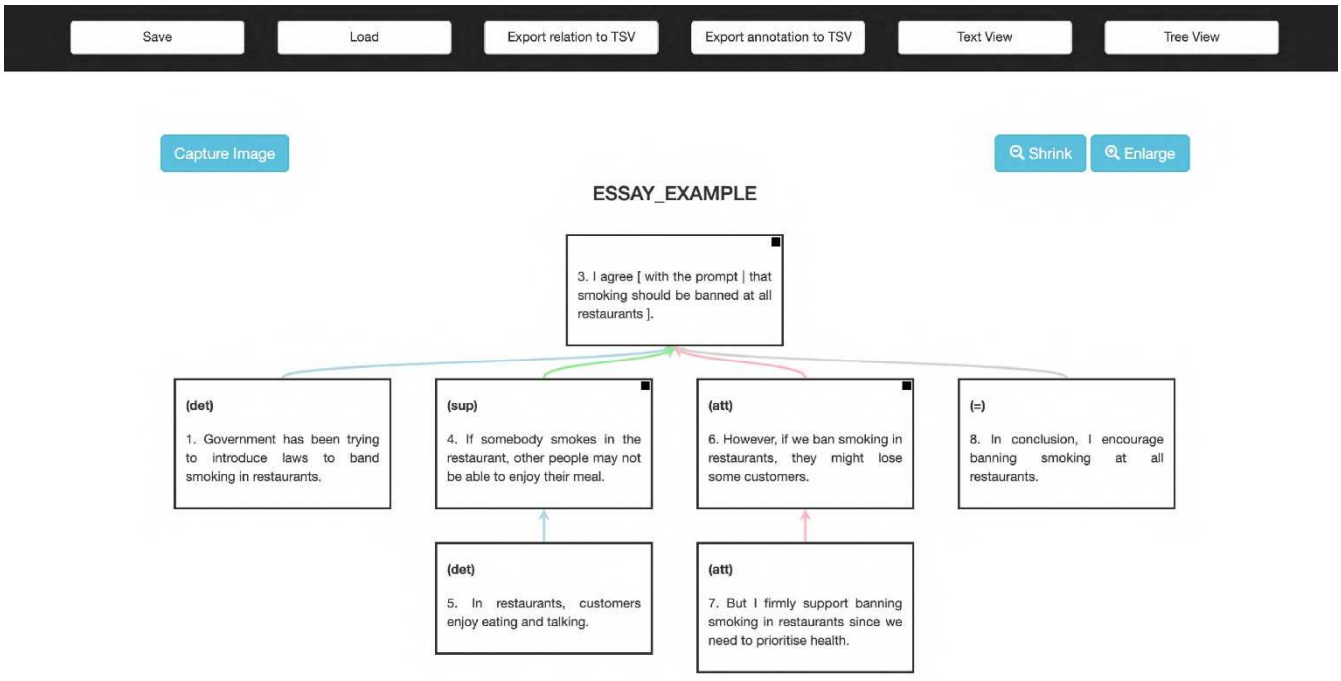

図 3　TIARA の木構造ビュー

TIARA はテキストを文の並びとして見せるテキストビュー(図 2)と文関係の関係を中心に見せる木構造ビュー (図 3)の 2 つのモードを持ち、2 つのビューを適宜切り替えることによって、テキスト構造のアノテーションが効率よくおこなえるように設計されている。テキストビューでは文を含む箱をドラッグ&ドロップすることによって文順序を簡単に変更できる。

3.2 テキスト構造の解析手法

テキストにおける文間の関係を同定することは、テキストの一貫性を向上させる上での第一歩である。作成したコーパス ICNALE-AS2R を訓練データとして、ニューラルネットを用いて文間の関係を自動同定する手法を開発した。提案手法は各文が関係付けられる文を同定する文リンキングとその関係種別を同定する関係同定の 2 つの処理からなる。文リンキングについては、各文が関係付けられる文との距離を各文に付与する系列ラベリングとして定式化する方法と文の依存構造解析の手法を文間の依存関係に拡張して適用する手法を試みた。関係同定については、文間の関係をコーパスに付与されている支持、攻撃、詳細化、再言明に分類する課題として定式化し、ニューラルネットを用いて微調整をおこなうモデルとおこなわないモデルを試した。

文リンキングについては、文の表現として、文を BERT モデル[6]で単語ごとにエンコードしそれを集約して文の表現を作る方法と SBERT モデル[7]で直接文の表現を作る方法を試みた。解析モデルについては系列ラベリングと依存構造解析を選択肢とした。これらの文表現と解析モデルの組合せで実験をおこなった結果、文を SBERT によって表現し、依存構造解析を使う手法がもっとも性能がよかった。系列ラベリングの手法では、テキストの議論構造に関係ない文を選別するタスクを補助タスクとしてマルチタスク学習も試したが、際立った効果は得られなかった。また、関係同定では、BERT モデルを用いて微調整したモデルが最も性能が高かった。

3.3 テキスト構造を考慮した文順序の整列手法

文の順序を整列する課題については従来から研究されていたが、従来は文の集合を整列する問題であったのに対し、本研究が対象とするのは既にテキストとし

言語処理技術を利用したテキストの一貫性向上のための推敲支援

Revision support for improving text coherence by using language processing technology

ての体裁を持った文の並びを一部並び替えてテキスト全体の構成をさらによくする課題であり、より困難な課題となっている。提案手法は、図 4 に示すとおり (0) テキスト構造解析、(1) 談話構造解析で関係付けられた文対の局所的順序の決定、(2) 文の全順序の決定の 3 つの段階から構成される。ICNALE-AS2R を用いた実験をおこない、言語生成の分野で提案されている Maximum Local Coherence[8]モデルや文の並び替え課題で最高性能を報告している Topological Sorting モデル[9]と比較して提案手法が著しく性能が高かった。しかしながら、提案手法の出力は改善前の元テキストを越えることはできなかったことから、エラー分析をおこなった結果、議論構造解析の精度が隘路となっていることを明らかなった。結論として、テキスト構造解析の精度をさらに高めることによって、元テキストの文順序を自動的に並び換えてテキストの一貫性を改善できる可能性があることを示した。

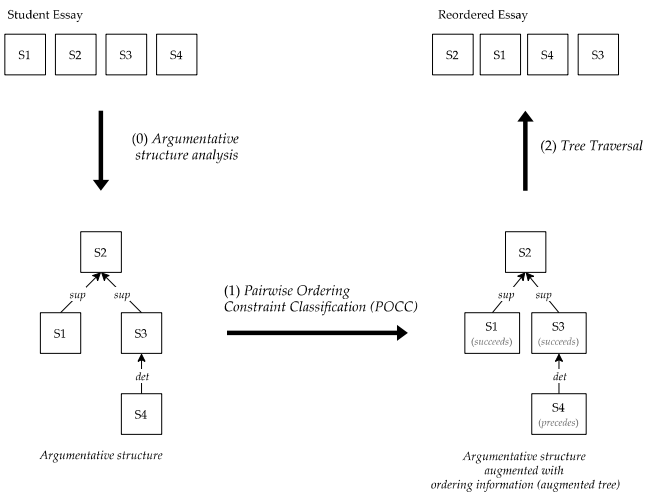

図 4　文整列手法の概要

4．将来展望

「よい」テキストとは何を厳密に定義することは難しい。スペルミスや文法的な誤りがないことは大前提であるが、文法的に正しい文の集合が必ずしも「よい」テキストになるとは限らない。テキストにはその内容を想定する読者に伝える目的があり、この目的を達成するためには、その構成要素である個々の文が主張する内容を互いに関連付け、適切な論理構造を組み上げる必要がある。もうひとつの側面として、テキストの読み易さがある。連続する文が意味的に関連付けられて話題が自然に推移するテキストはその読み易さに寄与する。本研究では前者をテキスト中のテキスト構造として、後者をテキストの一貫性としてとらえ、学生の書いたエッセイに議論構造と一貫性を向上させるための修正記録をアノテーションしたコーパスを作成した。テキスト構造をエッセイにアノテーションしたコーパスは、Iryna Grevych (Darmstadt 大学)、Manfred Stede (Potsdam 大学)、Vincent Ng (Texus Dallas 大学)らのグループが作成しているが、テキスト構造に加えて一貫性を向上させる修正をアノテーションしたコーパスを作成した試みはなく、このコーパスはテキストの議論構造と一貫性の関係を研究する上での有用な基礎データとして利用できる。この分野の研究に貢献するために、本研究で構築したコーパス(ICNALE-AS2R)は特定非営利活動法人言語資源協会(https://www.gsk.or.jp)から公開(GSK2021-A)している。また、TIARA はソースコードと共に公開している(https://github.com/wiragotama/TIARA-annotationTool)。

現行の日本の教育システムには、論理的で一貫性のあるテキストを書く能力を涵養するカリキュラムが整っているとは言いがたい。本研究が目的としているテキストの構造レベルの推敲支援が実現できれば、作文演習などに利用することによって教育的な貢献も期待できる。

参考文献

[1] Grammarly: https://www.grammarly.com

[2] Ginger: http://www.getginger.jp

[3] Robert Dale, Ilya Anisimoff, and George Narroway. HOO 2012: A report on the preposition and determiner error correction shared task. Proceedings of the Seventh Workshop on Building Educational Applications using NLP, pp. 54-62, 2012.

[4] Michael. A. K. Halliday and Ruquaiya Hassan. *Cohesion in English*. Longman, 1976.

[5] Shinichiro Ishikawa. The ICNALE Edited Essays: A dataset for analysis of L2 English learner essays based on a new integrative viewpoint. *English Corpus Linguistics*, 25, 1-14, 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171-4186, 2019.

[7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982-3992, 2019.

[8] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, Benno Stein. Computational argumentation synthesis as a language modeling task. *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 54–64,2019.

[9] Prabhumoye, Shrimai, Ruslan Salakhutdinov and Alan W Black. Topological sort for sentence ordering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2783–2792, 2019.

関連文献

(1) Jan Wira Gotama Putra, Kana Matsumura, Simone Teufel, Takenobu Tokunaga, TIARA 2.0: An interactive tool for annotating discourse structure and text improvement, *Language Resources and Evaluation*, pp. 1-25, 2021. Nov. DOI: 10.1007/s10579-021-09566-0

(2) Jan Wira Gotama Putra, Simone Teufel and Takenobu Tokunaga, Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance, *Proceedings of the 8th Workshop on Argument Mining*, pp. 12-23, 2021. Nov.

(3) Jan Wira Gotama Putra, Simone Teufel and Takenobu Tokunaga, Annotating argumentative structure in English-as-a-Foreign-Language learner essays, *Natural Language Engineering*, online 2021. Aug. DOI:10.1017/S1351324921000218

(4) Jan Wira Gotama Putra, Simone Teufel and Takenobu Tokunaga. Parsing argumentative structure in English-as-foreign-language essays, *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 97-109, 2021. Apr.

(5) Jan Wira Gotama Putra, Simone Teufel, Kana Matsumura and Takenobu Tokunaga, TIARA: A Tool for Annotating Discourse Relations and Sentence Reordering, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020),* pp.6912-6920, 2020, May.

この研究は、平成３０年度ＳＣＡＴ研究助成の対象として採用され、令和元年度～令和３年度に実施されたものです。