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

令和　５ 年 １０ 月 ３０ 日現在

機関番号：２１６０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ００８５０

研究課題名（和文）Feature visualizer and detector for scientific texts

研究課題名（英文）Feature visualizer and detector for scientific texts

研究代表者
ＢＬＡＫＥ　Ｊｏｈｎ（Blake, John）

会津大学・コンピュータ理工学部・上級准教授

研究者番号：８０６３５９５４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：言語特徴の使用パターンを可視化するツールを2つ開発しました。最初のツール(Feature visualizer)は、submitを使用したテキストを分析するものである。2つ目のツール（Feature detector）は、データベース内のテキストを分析するものである。両ツールとも、時制や結束などの言語特徴に注釈を付け、色付けする。

研究成果の学術的意義や社会的意義
The primary aim of this project is to develop an online resource that could assist Japanese writers of short research articles in the field of computer science to understand the prototypical generic features in such articles. This is envisaged to help them climb the cline of competence more quickly.

研究成果の概要（英文）：We have developed Feature Detection and Feature Visualization tools. The Feature Visualization tool comprises an annotated dataset of short research articles and a bank of multimodal materials which are displayed in the user interface of the Feature Visualizer. Here users can visualize particular rhetorical or language aspects, e.g. modality, tense and cohesion. Users then have the option to display additional multimodal explanations to understand the specific rhetorical or language features. In addition, two Feature Detection tools were created that can process student-submitted work. The first colorizes finite verb phrases according to one of twelve pedagogic tenses. The main feature detection tool enables users to gain feedback on deep grammatical features, namely information structure. The end weight, the information focus and information flow are automatically annotated, helping learners differentiate between unmarked, highly frequent usage and marked, rare usage.

研究分野：natural language processing

キーワード：scientific writing　genre awareness raising　rhetorical features　language features　information structure　pedagogic tool

１．研究開始当初の背景

Novice writers of scientific texts have access to specialist tools that can assist the drafting and editing processes, such as the Scientific Writing Assistant (Kinnunen et al., 2012) Academic Word Suggestion Machine (Mizumoto, 2017) and a corpus-based error detector (Blake, 2018). However, without sufficient exposure to a genre, it is difficult to write in the style expected by the community of practice (Lave and Wenger, 1991; Hyland, 2012). Students nowadays invariably search the Internet for advice. There are a few tools detecting specific language features. However, their focus is limited and little consideration has been given to their usefulness as learning tools. There is currently no interactive online tool that aims to help learners understand generic conventions in scientific writing. This feature visualizer and detector, therefore, will be the first online tool to enable students to understand generic conventions through discovery learning. Users expect online learning resources to be interactive, highly visual and multimodal (Hafner, Chik and Jones, 2015) and so video, audio and images will be harnessed within the interactive tool.

２．研究の目的

This research aims to discover to what extent can an online tool detect, visualize and enhance knowledge of generic integrity (Bhatia, 1999) in the scientific writing of computer science students. To do so it is first necessary to create the online tool.

３．研究の方法

The research method comprises two main stages: development and evaluation. The bulk of the research is concerned with the development and improvement of the system. The evaluation stage within the scope of this project is limited to student survey regarding perceptions in the change in awareness of generic conventions.

In the first year of the project we annotated a small corpus of short research articles that will form the dataset of the feature visualizer. We have also created a number of explanatory videos to be displayed in the online feature detector. We created some low-fidelity and high-fidelity prototypes in order to select a user-friendly interface with the required functionalities. The base for the feature visualizer was created using Django and Vue.js and deployed online.

In the second year, we created programs that match pre-annotated segments of texts. These programs that run on raw text. We improve the feature detector by integrating more functionalities, such as tense-aspect identification and various types of information structure. The tense-aspect identification function classifies and labels grammatical tenses using the twelve commonly-used terms (e.g. past progressive, future perfect, etc.). The tense-aspect identification function also classifies finite verbs by voice, and so that feature will also be available for users. The information structure function, which identifies information focus, information flow and end-weight was deployed.

In the final year, we refined the algorithms used in various functions. We added sophisticated natural language pipelines to automatically highlight cohesive features and show coherence using entity detection.

４．研究成果

We have developed Feature Detection and Feature Visualization tools as described in the initial proposal.

Feature visualizer

The Feature Visualization tool comprises an annotated dataset of short research articles and a bank of multimodal materials which are displayed in the user interface of the Feature Visualizer. Here users can access research articles categorized by four categories (applied, empirical, experimental and theoretical), and then within each article, they can visualize particular rhetorical or language aspects, e.g. modality, voice and tense (See Fig. 1). Users then have the

option to display additional multimodal explanations to understand the specific rhetorical or language features.

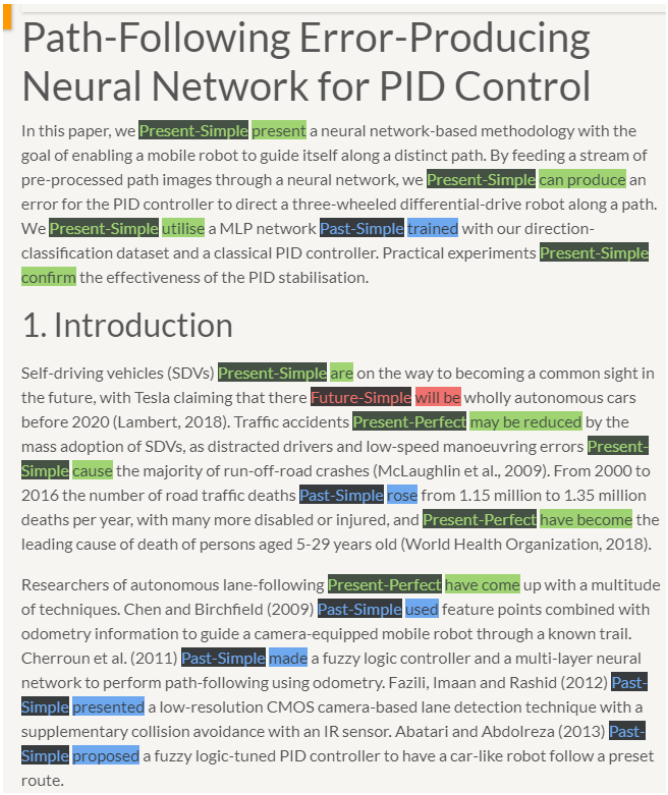

Fig 1. Screenshot of Feature Visualizer with tenses highlighted.

Two particularly challenging features to develop were the automatic cohesion and coherence features. Cohesion within each paragraph is shown by colorizing the entities that are repeated, which is achieved by linking anaphoric references to the antecedents. Coherence is shown at paragraph level by a summary noun phrase, usually a single noun.

Other technical challenges overcome in this project include displaying fully formatted texts in rich texts. This was particularly onerous as each article was initially sources as a PDF. The images, tables, equations and text had to be separated. The text had to be annotated to show the organizational structure, namely sections and rhetorical moves as the current state-of-the-art natural language processing pipelines were unable to automatically identify these with the required degree of accuracy for pedagogic materials.

Feature detection tools

In addition to the Feature Visualizer, two Feature Detection tools were created that can process student-submitted work.

Feature detector 1: Tense identifier

The first tool shown in Fig 2 focuses on tenses. This was separated into a discrete tool, given its applicability to multiple user groups. The tool colorizes finite verb phrases according to one of twelve pedagogic tenses (i.e. the commonly taught forms, e.g. present perfect progressive). Additional information (e.g. voice, verb sense, and dictionary definition) regarding the verb phrase is available by placing the cursor over the colorized finite verb.

Feature detector 2: Integrated tool (Language Feature Detector)

The integrated feature detection tool enables users to gain feedback on multiple language features. Commonly used features, such as automatic identification of readability using a range statistical methods, and text profiling using a variety of vocabulary lists are available.
The text profiling feature colorizes the text by the vocabulary categories within each list. In

addition, the dictionary definitions of computer science related words can be accessed directly via WordNet. Lesser annotated aspects namely deep grammatical organization principles of information structure. Information structure principles of end weight, the information focus and information flow are automatically annotated, helping learners differentiate between unmarked, highly frequent usage and marked, rare usage. Fig. 3 shows the end weight statistics calculated for a simple text.

Fig. 2 Screenshot of output from Tense Identifier

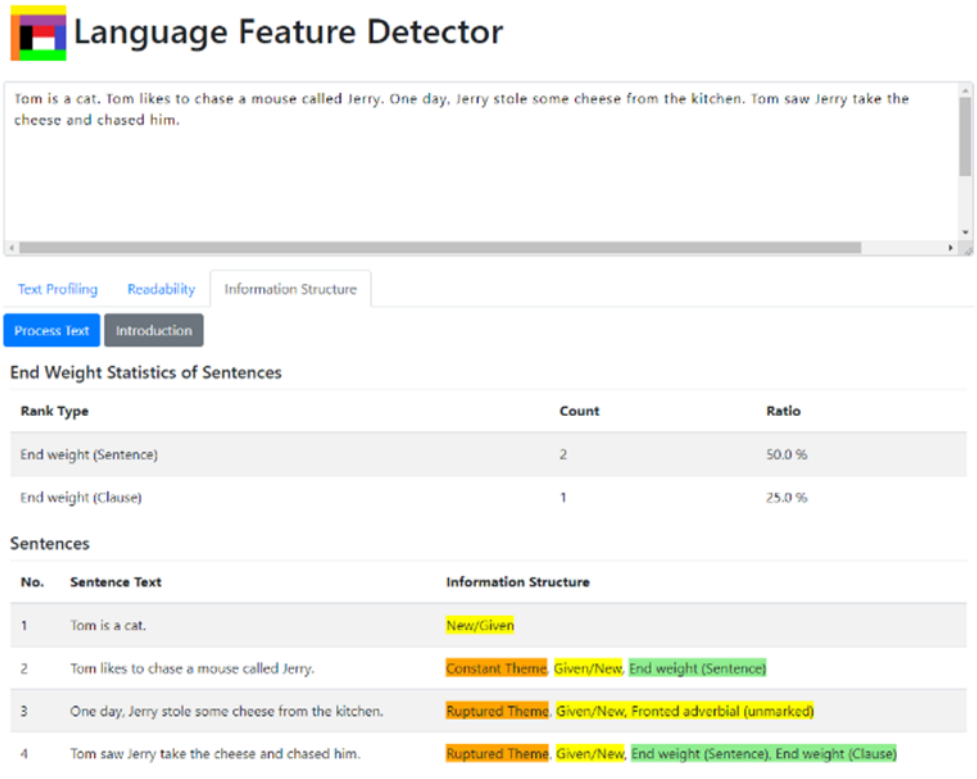

Fig. 2 Screenshot of output from Language Feature Detector

References

Bhatia, V. (1999). Generic integrity in document design. Document Design, 1 (3), 151-163.

Blake, J. (2018). Corpus-based error detector for Computer Science. In Y. Tono and Isahara, H. (Eds.), Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference, (pp.50-54). Takamatsu, Japan.

Hafner, C. A., Chik, A., & Jones, R. H. (2015). Digital literacies and language learning. Language Learning & Technology, 19 (3), 1–7.

Hyland, K. (2012). Disciplinary Identifies: Individuality and Community in Academic Discourse. Cambridge: Cambridge University Press.

Lave, J. and Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge: Cambridge University Press.

Kinnunen et al., (2012). Scientific writing assistant. http://cs.joensuu.fi/swan/

Mizumoto, A. (2017). Academic Word Suggestion Machine. http://langtest.jp/awsum/

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　5件／うちオープンアクセス　5件）

１．著者名 Blake, John	４．巻 1
２．論文標題 Development of an online tense and aspect identifier for English	５．発行年 2020年
３．雑誌名 CALL for widening participation: short papers from EUROCALL 2020	６．最初と最後の頁 36--41
掲載論文のDOI（デジタルオブジェクト識別子） 10.14705/rpnet.2020.48.1161	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Blake, John	４．巻 1
２．論文標題 English Verb Analyzer: Identifying tense, voice, aspect, sense and grammatical meaning in context for pedagogic purposes.	５．発行年 2020年
３．雑誌名 Proceedings of 8th Swedish Language Technology Conference 2020	６．最初と最後の頁 1--5
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Blake, John	４．巻 2
２．論文標題 Automatic identification of tense and grammatical meaning in context	５．発行年 2020年
３．雑誌名 Proceedings of the International Conference on Computers in Education 2020	６．最初と最後の頁 739--742
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Blake, John	４．巻 1
２．論文標題 Generic integrity: Visualizing lexicogrammatical features in scientific articles.	５．発行年 2020年
３．雑誌名 Selected online proceedings of the British Association of Applied Linguists Annual Conference 2019	６．最初と最後の頁 1--3
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Blake, John	4．巻 1
2．論文標題 Annotated scientific text visualizer: Design, development and deployment	5．発行年 2019年
3．雑誌名 CALL and complexity - EUROCALL	6．最初と最後の頁 45-50
掲載論文のDOI（デジタルオブジェクト識別子） 10.14705/rpnet.2019.38.984	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　3件）

1．発表者名 Blake, J.
2．発表標題 Multimodal content creation for pedagogic purposes: Lessons learned
3．学会等名 GLoCALL 2021（国際学会）
4．発表年 2021年～2022年

1．発表者名 Blake, J.
2．発表標題 Automatic annotation of information structure
3．学会等名 Contrast and Annotation IS 2021: International Workshop on the Expression of Contrast and the Annotation of Information Structure in Corpora
4．発表年 2021年～2022年

1．発表者名 Blake, J.
2．発表標題 Detecting focus, flow and end weight in research articles
3．学会等名 6th International Conference of Asia-Pacific LSP and Professional Communication Association
4．発表年 2021年～2022年

１．発表者名 Blake, John
２．発表標題 Generic integrity: Visualizing lexicogrammatical features in scientific articles
３．学会等名 British Association of Applied Linguistics Conference（国際学会）
４．発表年 2019年～2022年

１．発表者名 Blake, J., Pyshkin, E. and Pavlic, S.
２．発表標題 Automatic detection and visualization of information structure in English
３．学会等名 Natural Language Processing and Information Retrieval（国際学会）
４．発表年 2022年～2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	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
研究分担者	Ｍｏｚｇｏｖｏｙ　Ｍａｘｉｍ (Mozgovoy Maxim) (60571776)	会津大学・コンピュータ理工学部・准教授 (21602)	

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