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　元 年　 ６ 月　 ３ 日現在

機関番号：１１１０１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ０２４６９

研究課題名（和文）統辞・意味解析情報タグ付き日本語ツリーバンクからの視覚意味情報の抽出と応用

研究課題名（英文）The Extraction and Application of Visual and Semantic Information from Japanese Treebanks with Syntactic and Semantic Analysis Annotation

研究代表者

バトラー　アラステア（Butler, Alastair）

弘前大学・人文社会科学部・准教授

研究者番号：９０５８８８７３

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究の目的は日本語と英語の分析に基づき、意味情報、特に、接続詞、述語、項、演算子束縛に関する情報を視覚化しアクセス可能にする手法を開発することであった。

本研究の重要な成果は、解析済みのコーパスデータから読み取られた意味関係を視覚的に表示するツールの開発である。このツールによって、例えば、意味役割、照応関係などの、文中あるいは談話中に見られる多くの関係を描くことができる。また、このツールを使うことで、コーパスのアノテーターは自身の文（あるいは談話）の解釈がアノテーションによって適切に捉えられているか否かを確認することができるようになる。

研究成果の学術的意義や社会的意義
視覚化ツールは次の解析情報付きコーパスの構築で利用されている：(1) 英語（TSPC, 7026ツリー, 87182語, http://www.compling.jp/ajb129/tspc.html）, (2) 現代日本語（NPCMJ, 30460ツリー, 505319語, http://npcmj.ninjal.ac.jp/interfaces), (3) 上代日本語（MYS97, 159ツリー, 2549語, http://www.compling.jp/mys97）。これにより, 以前は不可能であった依存関係の層を視覚的に示すということが、数千の実例という規模で可能になった。

研究成果の概要（英文）：The research aim has been to develop methods of visualising and making accessible semantic information from analyses of Japanese and English, notably: connective, predicate, argument, and operator-binding information.

The key result of this work has been the development of a visualisation tool for semantic relationships derivable from parsed corpus data. This tool has the capability to present many relationships found internally to sentences and across discourse, e.g., providing a way to visualise semantic roles, and anaphoric dependencies. This aids human annotators building a parsed corpus to assess whether their interpretations of sentences (and discourse) have been adequately captured by the annotation.

研究分野：linguistics

キーワード：semantic dependencies　parsed corpus　visualisation　annotation　predicate arguments　discourse relations

１．研究開始当初の背景

This work began as a continuation of research into the automatic analysis of natural language, focusing on semantic relations in Japanese and English. The overarching aim has been to establish a new method of corpus construction that involves providing texts with enough analysis information to thereafter be able to reliably support an automatic calculation of derived information. Derived information includes, for example, establishing scope information, resolving anaphoric relations, and producing detailed valence patterns. Equally important, there has been the aim to apply the method with the actual development of substantial corpora for Japanese and English, that is, collections of texts, with tens of thousands of sentences, that are given analysis following the method. In order to make it possible to keep track of the results, a notable aim was to develop a new visual presentation method applicable to the representation of the logical meaning of sentences. In this regard, research was conducted on two tasks: displaying familiar “dependencies” performed in university education, and visualizing semantic analysis information using Flame Graphs. The latter is a technique for visualizing aggregations of results gathered from across many instantiations of relatable phenomena, such as contexts of occurrence.

２．研究の目的

A key research purpose was to develop methods of visualizing and making accessible semantic information from analyses of Japanese and English, notably: connective, predicate, argument, and operator-binding information. Such information enables, for example, the mapping out of language binding dependencies, which has proved particularly relevant as a method to reconstruct unpronounced argument information (zero pronouns) for Japanese. Another major strength of this work has been the ability it gives to extract valence patterns for predicates, an essential part of word meaning.

The work on visualization has focused on ways to preserve descriptive content. Innovations include: adding coded indexing carrying ontological/sense information, and folding represented material around “inverse roles” to compact hierarchical structure. Presence of indexing can clutter visualization, but it also makes possible flipping between different views of content, e.g., with a dimension to capture sentence content, and a dimension for an overview of total discourse content. In contrast to indexing, the addition of inverse roles simplifies visualization thorough the normalization of hierarchical information. As a side effect, this created the foundation for supporting a method of natural language generation, that is, a way to get back to natural language from a semantic representation.

The work on developing methods of visualizing and making accessible semantic information also focused on ways to embed information back into parsed data. This led to the enrichment of the existing corpus data with a second layer of special-purpose annotation made up of indexing information. This corpus semantic information can now be searched because of a transformation to an XML format, TIGER-XML, that includes a structure sharing mechanism (multi-dominance) that can be queried.

３．研究の方法

To carry out the corpus creation, it was necessary to continue developing a method for reaching semantic representations automatically from syntactic parsed representations, and to create a large base of already analyzed and human checked syntactic structures that can be transformed to semantic representations. The establishment of such a base formed training data for creating yet more like data. During the course of the project the range of data analyzed was continuously extended to more genres, including to historical Japanese texts.

Throughout the project, the semantic component continued to be developed, especially in its use as a basis for visualizing dependencies. In this regard an indexing component was extended so as to produce the character-indexed report format of FrameNet. This allowed creation of browsable reports that display semantic dependencies in a very intuitive way.

The developed visualization revealed layers of dependencies that were previously hidden. At the same time, the tool revealed inadequacies of analyses in the state of the corpus data. Throughout the project, there has been sustained investment of time and energy into improving the quality of what was already annotated data. Steadily the project has evolved through continuing cycles of improving the existing annotation, with improvements magnifying what is possible with the same dataset.

Finally, effort was placed into publicizing the results of the project domestically and abroad at academic conferences.

４．研究成果

The key result of this project has been the development of a visualization tool for semantic relationships derivable from parsed corpus data. This tool has the capability to visually present many relationships found internally to sentences and across discourse, e.g., providing a way to visually explore semantic roles, and anaphoric dependencies. This assists human annotators building a parsed corpus to assess whether their interpretations of sentences in discourse contexts have been adequately captured by the annotation.

The visualization tool has been used in the creation of annotated corpora for Contemporary English (TSPC; 7026 trees; 87182 words; http://www.compling.jp/ajb129/tspc.html), Contemporary Japanese (Keyaki Treebank; 51617 trees, 822151 words, 1489934 characters; http://www.compling.jp/keyaki/, and NPCMJ; 30460 trees, 505319 words; http://npcmj.ninjal.ac.jp/), and Old Japanese (MYS97; 159 trees; 2549 words; http://www.compling.jp/mys97). This has revealed layers of dependencies that were not easily visible before, and certainly not available on a corpus scale, with thousands of attestations to explore (that is, both search and see).

Research results produced by this project can also now be seen in the interfaces of the NINJAL Parsed Corpus of Modern Japanese (NPCMJ; http://npcmj.ninjal.ac.jp/interfaces/), where, aside from a default tree view of the syntactic annotation, examples of the corpus can be seen as predicate logic formulas capturing semantic content (semantic view), as well as a view that embeds the calculated semantic content into the trees as indexing information (indexed view).

５．主な発表論文等

〔雑誌論文〕（計　14 件）

①竹内孔一，バトラー アラステア，長崎郁，ホーンスティーブンライト，2019．PropBank スタイルの意味役割タグを導入した述語項構造シソーラスと NPCMJ への付与計画．Proceedings of the Twenty Fifth Annual Meeting of the Association of Natural Language Processing. pages 136-138．査読有

②Alastair Butler and Stephen Wright Horn. 2018. Parsed Annotation with Semantic Calculation. Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories. pages 39-51. 査読有

③Horn Stephen Wright, Alastair Butler, Iku Nagasaki and Kei Yoshimoto. 2018. Derived mappings for FrameNet construction from a parsed corpus of Japanese. LREC 2018 Proceedings. pages 28-32. 査読有

④長崎郁，アラステア・バトラー，スティーブン・ライト・ホーン，プラシャント・パルデシ and 吉本．2018．統語解析情報付きコーパス検索用インタフェースの開発．『言語処理学会第 24 回年次大会発表論文集』．pages 1123-1126．査読無

⑤Horn Stephen Wright, Alastair Butler and Kei Yoshimoto. 2017. Keyaki Treebank segmentation and part-of-speech labelling. 『言語処理学会第 23 回年次大会発表論文集』．pages 414-417．査読無

⑥Alastair Butler. 2016. From meaning representations to syntactic trees. Proceedings of the Thirteenth International Workshop of Logic and Engineering of Natural Language Semantics. pages 147-160．査読有

⑦Alastair Butler. 2016. DynamicPower at SemEval-2016 Task 8: Processing syntactic parse trees with a Dynamic Semantics core. Proceedings of SemEval-2016. pages 1148-1153. 査読有

⑧Alastair Butler. 2016. Deterministic natural language generation from meaning representations for machine translation. Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation. pages 1-9. 査読有

⑨Alastair Butler, 吉本啓, 岸本秀樹 and プラシャント・パルデシ. 2016. 「統語・意味解析情報付き日本語コーパスのアノテーション」. 『言語処理学会第 22 回年次大会発表論文集』. pages 589-592. 査読無

⑩Alastair Butler. 2015. Something, namely "SomethingNamely". Proceedings of the Twelfth International Workshop of Logic and Engineering of Natural Language Semantics. pages 215-228. 査読有

⑪Alastair Butler. 2015. Round trips with meaning stopovers. Proceedings of the 1st Workshop on Semantics-Driven Statistical Machine Translation. pages 1-10. 査読有

⑫Pardeshi Prashant, Alastair Butler, Kei Yoshimoto and Hideki Kishimoto. 2015. 「統語・意味解析情報付き日本語コーパスの開発」. 『言語処理学会第 21 回年次大会発表論文集』. pages 20-23. 査読無

⑬Alastair Butler and Kei Yoshimoto. 2015. Large scale semantic representation with flame graphs. 『言語処理学会第 21 回年次大会発表論文集』. pages 301-304. 査読無

⑭Alastair Butler, Shota Hiayama and Kei Yoshimoto. 2015. Coindexed null elements for a Japanese parsed corpus. 『言語処理学会第 21 回年次大会発表論文集』. pages 708-711. 査読無

〔学会発表〕(計 23 件)

①竹内孔一, Alastair Butler, 長崎郁, スティーブン・ライト・ホーン. PropBank スタイルの意味役割タグを導入した述語項構造シソーラスと NPCMJ への付与計画.言語処理学会 第 25 回年次大会. 2019 年

②With Iku Nakasaki, Susanne Miyata and Alastair Butler. Changing the morphological base of the NPCMJ. Collaborative Research Project Meeting. 2019.

③Alastair Butler, Stephen Wright Horn. Parsed annotation with semantic calculation. 17th International Workshop on Treebanks and Linguistic Theory. 2018.

④Alastair Butler. A unified interface for exploring English and Japanese. 日本英語学会第 36 回大会シンポジウム. 2018 年.

⑤Alastair Butler, Stephen Wright Horn. English/Japanese contrastive study based on normalisation, a step in the semantic processing.日本英語学会第 36 回大会シンポジウム. 2018 年.

⑥スティーブン・ライト・ホーン, 鴻野知暁, アラステア・バトラー, 小木曽智信, フレレスビッグ・ビャーケ.「「オックスフォード・NINJAL 上代コーパス」の公開」. 日本語学会 2018 年秋季大会. 2018 年.

⑦Alastair Butler, Stephen Wright Horn. Tools and practices for annotating discourse.「統語・意味解析コーパスの開発と言語研究」研究発表会. 2018 年.

⑧Alastair Butler, Stephen Wright Horn, Iku Nagasaki and Kei Yoshimoto. Derived mappings for FrameNet construction from a parsed corpus of Japanese. LREC 2018. 2018.

⑨Alastair Butler, 長崎郁, スティーブン・ライト・ホーン, プラシャント・パルデシ, 吉本啓.「統語解析情報付きコーパス検索インタフェースの開発」. 言語処理学会第 24 回年次大会 (NLP2018). 2018 年.

⑩Alastair Butler, Stephen Wright Horn and Iku Nagasaki. Parsed corpus (ad)ventures. 共同研究発表会「統語・意味解析コーパスの開発と言語研究」. 2017 年.

⑪Alastair Butler, Susanne Miyata. Developing a model of typical Japanese grammar development: The role of parsed corpora and parsing programs. Exploiting Parsed Corpora: Applications in

Research, Pedagogy, and Processing. 2017.

⑫Alastair Butler, Stephen Wright Horn and Iku Nagasaki. Seeding lexical semantics: resources using parsed corpora. Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing. 2017.

⑬Alastair Butler, Stephen Wright Horn. Annotating syntax and lexical semantics with(out) indexing. Logic and Engineering of Natural Language Semantics (LENLS 14). 2016.

⑭Alastair Butler, Ai Kubota, Shota Hiyama and Kei Yoshimoto. Treebank annotation of FraCaS and JSeM. Logic and Engineering of Natural Language Semantics (LENLS 13). 2016.

⑮Alastair Butler. From meaning representations to syntactic trees. Logic and Engineering of Natural Language Semantics (LENLS 13). 2016.

⑯ Alastair Butler. Parsed Corpus Semantics. New Landscapes in Theoretical Computational Linguistics. 2016 (Invited Speaker).

⑰Alastair Butler. Deterministic natural language generation from meaning representations for machine translation. 2nd Workshop on Semantics-Driven Machine Translation. 2016.

⑱ Alastair Butler. Something, namely “SomethingNamely”. Logic and Engineering of Natural Language Semantics (LENLS 12). 2015.

⑲Alastair Butler. Processing linguistic expressions to reach semantic content for application. 『日本ソフトウェア科学会第 32 回大会』 (JSSST2015). 2015 (Invited Speaker).

⑳Alastair Butler. Round trips with meaning stopovers. Semantics-Driven Statistical Machine Translation: Theory and Practice (S2MT), Workshop in conjunction with ACL 2015. 2015.

㉑Alastair Butler, Kei Yoshimoto. Development of Japanese Corpus Tagged with Syntactic and Semantic Information. The 18th Joint Workshop on Linguistics and Language Processing, Theme: Formalism and Functionalism in Linguistics. 2015.

㉒Alastair Butler, Kei Yoshimoto. Large scale semantic representation with flame graphs. The Twenty First Annual Meeting of the Association for Natural Language Processing (NLP2015). 2015.

㉓Alastair Butler, Shota Hiyama and Kei Yoshimoto. Coindexed null elements for a Japanese parsed corpus. The Twenty First Annual Meeting of the Association for Natural Language Processing (NLP2015). 2015.

〔図書〕（計　1 件）
①Alastair Butler. 2015. Linguistic Expressions and Semantic Processing: A Practical Approach. Springer-Verlag. 172 pp.

〔その他〕
ホームページ等

①Treebank Semantics
http://www.compling.jp/ajb129/ts.html

Treebank Semantics implements obtaining meaning representations from parsed corpora.

②The Treebank Semantics Parsed Corpus (TSPC)
http://www.compling.jp/ajb129/tspc.html

This is a corpus resource for English, with annotation for 7026 trees, 87182 words. Highlights include: bracketed constituent structure, assignments of grammatical role and function, and binding information (to resolve anaphoric dependencies). The parsed data, and further results of analysis (e.g., derived indexing, word dependencies, generated semantic representations), are made accessible through a web-based interface.

③The Keyaki Treebank Homepage
http://www.compling.jp/keyaki/

This is a corpus resource for Japanese, with annotation for 51617 trees, 822151 words, 1489934 characters. Highlights include: bracketed constituent structure, assignments of grammatical role and function, and zero elements.

④The NINJAL Parsed Corpus of Modern Japanese (NPCMJ)
http://npcmj.ninjal.ac.jp/interfaces/

This is a corpus resource for Japanese, with annotation for 30460 trees, 505319 words. This is a continuation and extension/refinement of the freely redistributable parts of the Keyaki Treebank. The parsed data, and further results of analysis (e.g., derived indexing, word dependencies, generated semantic representations), are made accessible through a web-based interface.

⑤The Man'yoshu97 Parsed Corpus (MYS97)
http://www.compling.jp/mys97/

This is a corpus resource that provides detailed parsed annotation for the first 97 poems of the Man'yōshū, which contains the oldest attested forms of the Japanese language, Old Japanese.

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：