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　４ 年　 ６ 月　 ６ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ００５５２

研究課題名（和文）コンコーダンサや言語処理で利用可能な蓄積型の名詞・述語項構造辞書の構築

研究課題名（英文）Construction of Japanese Predicate-Argument Structure Dictionary for Natural Language Processing and Linguistic Analysis with Concordancer

研究代表者

竹内　孔一（Takeuchi, Koichi）

岡山大学・自然科学学域・准教授

研究者番号：８０３１１１７４

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：述語項構造辞書の拡張として新たにPropBank形式の意味役割体系を導入して日本語に適用した体系を整備した.具体的にはArg0, Arg1やArgMのように必須項と付加詞を分けて記述して使役態や受動態，被害受身など異なる構文にも項が追跡できるように拡張した.また辞書項目を拡張・整備しWebサイトを通じて公開した．述語項構造をテキストに付与するモデルとして深層学習モデルやベイズの定理に基づくモデルを適用し精度が向上する手法を明らかにした．意味役割付与システムを利用して，テキストに対してブロックベースでパターンを組み合わせて意味役割と概念フレームを利用してテキストを検索するコンコーダンサを開発した．

研究成果の学術的意義や社会的意義
本研究成果の中心的な意義は，従来，言語学における語彙意味論の中で議論されてきた意味役割と概念フレームといった述語項構造を実際に工学的に作成し応用システムまで開発したことである．PropBank形式を体系に導入することで，日本語特有の構文である被害受身などを同じ枠組でどのように記述するかを明らかにして国際会議で発表した．また，日本語の述語項構造を利用したアプリケーションとしてコンコーダンサを実装したことは，語彙意味論の成果を使える形で利用できる新たな枠組を提案したことになる．言語学習者などが言語表現を探す際に利用できるシステムになっており，言語理論を社会で利用する一例を具体的に示した．

研究成果の概要（英文）：First, we incorporate the semantic role system defined in PropBank into the Predicate-argument structure Thesaurus. This extension allows us to capture the semantic relations between predicates and their arguments even in different constructions, such as the causative, passive and adversative passive forms, by describing both numbered arguments and named arguments (e.g., Agent and Theme). The dictionary entries have also been maintained and published on the website. Second, for construction of automatic annotation system of predicate-argument structures, we have applied various kinds of models such as deep learning models and models based on Bayes' theorems and revealed the approaches to improve the performance. Finally, we have developed a concordancer using an automatic annotation system of predicate-argument structures. In the concordancer, the users can combine block-based patterns that allow us to extract matched texts .

研究分野：自然言語処理

キーワード：コンコーダンサ　述語項構造　ブロックベース　意味役割　概念フレーム　パターンマッチ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)　英語に対して述語と項構造を付与した辞書とアノテーションデータ (PropBank) が構築されていた．一方，日本語の述語に対して述語項構造シソーラスが公開されていたが，名前の意味役割を中心に整理されていた．

(2) 述語項構造シソーラスの体系に基づく付与データとして国立国語研究所が作成したコーパス BCCWJ に付与されている事例が中心であった．BCCWJ の例文は作業者が付与例を選択したため受動態など態の違いによる事例はほとんど取り上げられていない状態であった．

(3)　述語項構造は述語と係り関係にある項に対して意味的な関係を付与する構造は言語学における語彙意味論から提案されていたが，一方で，言語処理としての利用は限られており，近年利用できるような形式での応用処理は提案されていなかった．

２．研究の目的

(1) 述語項構造シソーラスの記述内容を整備する．具体的な文中の表現に対して述語項構造シソーラスで定義されている概念フレームと意味役割を適用して，不足している概念フレームがあれば追加し整理する．また，能動態を中心に付与してきたため，使役や受動態など異なる態における意味役割の付与方法を明らかにする．

(2) 述語項構造シソーラスの自動拡張およびテキストに対する自動付与方法を明らかにする．

(3) 述語項構造とは述語の解釈である概念フレームと係り関係の意味的な関係を記述した意味役割から成る．このような意味的な関係をユーザが指定することで，例文から取り出すことができるコンコーダンサを作成する．例えば「X(対象)を購入する(概念 ID 71)」を指定すれば，「本を買う」「企業を買収する」など「買う」という概念でかつ，買われるもの(対象)の部分をテキストから取り出す．さらに，これらをユーザが簡単に組み合わせて，ユーザのテキストに適用できるようにブロック形式でパターンを組み合わせられるコンコーダンサを作成する．これにより，述語項構造を利用する環境を構築する．

３．研究の方法

(1)　新たな例文に対して述語項構造シソーラス辞書の概念フレームと意味役割を作業者が適用して付与できない事例を収集する．付与ができない事例に対して，辞書の概念フレームを見直して拡張する．拡張した辞書は Web 上で共有して更新後の体系で新たな事例を分析する．

(2)　機械学習を利用したモデルを適用して辞書拡張およびテキストに対する述語項構造の自動付与方法を検討する．学習データとして述語項構造シソーラスの事例の他に BCCWJ の付与例などを利用する．

(3)　各パターンをブロックに対応させて組み合わせで整合性のあるパターンを作成してテキストに対してパターンマッチを実行する機構を構築する．具体的には prolog による構造化を利用する．この際，マッチ結果表示に必要な項の設計，および意味役割付与システムとの対応を明らかにする．

(4) ブロック型プログラミング言語として JavaScript で動作する Blockly を利用したブラウザーベースの枠組を利用したコンコーダンサシステムを開発する．これにより，OS に依存せずに述語項構造を検索で利用できる環境を構築する．

４．研究成果

(1)　述語項構造シソーラスに対して基本表現として約 600 事例を追加し，約 2.4 万事例を公開した．さらに PropBank 形式の数字による意味役割 Arg0，Arg1 や ArgM を導入して全ての事例に付与を行い整備した．この際，使役や受動態，心理的な動詞に対する PropBank 形式の意味役割付与の方法を定義して会議で発表した(Takeuchi et al．2020 を参照)．心理動詞の場合，日本語には被害受身が存在するが，能動態から一貫して同じ意味役割になるように整理した．具体例を下記に示す．

［Arg1 愛は］［Arg0 真理の態度に］［Frame:苦しみ(243) 困る］
［Arg0 真理は］［Arg1 愛を］［Frame:苦しみ(243) 困らせる］
［Arg1 愛は］［ArgM-TMP 子供の頃から］［Arg0 真理に］［Frame:苦しみ(243) 困らされてばかりいる］

PropBank 形式の意味役割を導入することで大きく拡張した点として必須項と付加詞の明示が挙げられる．従来の名前による意味役割「動作主」「対象」では必須項か付加詞かの明示はされていなかったが，PropBank 形式を導入することで，必須項を明示するように変更した．また名前の意味役割は「使役」「動作主」などの違いを記述する一方で，PropBank 形式では Arg0 と縮約されてしまうため，従来の名前による意味役割と PropBank 形式の両方を付与する形式で構築した．

(2) 機械学習を利用してシソーラス拡張方法の提案，および意味役割付与における機械学習モデルについて実験を行い適切なモデルについて論文で発表した．

シソーラスの拡張では英語圏で開発されている AutoExtend を利用して日本語の WordNet に対して辞書構造を埋め込んだベクトルを作成し，学習した辞書にない未登録の日本語動詞に対して推定された synset が適切なものが多いことを明らかにした(Kou & Takeuchi 2019 を参照)．

また意味役割付与モデルの構築では Bidirectional で GRU (Gated Recurrent Unit)を利用した深層学習モデルを新たに提案して従来の機械学習モデルの手法に比べて正解率が 10%以上向上することを示した．さらに，図 1 に示すように学習データを 1/2 から 1/8 まで減少させた場合の意味役割付与精度について学習データが少ない場合でも提案する Bidirectional GRU モデルでは精度の減少がゆるやかであることを明らかにした．図 1 の破線のモデルは GDA データ(GSK2009-B)を転移学習として利用した場合である．全学習データを利用した場合の正解率は転移学習を利用しない方が高いが，学習データが少ない場合にはより正解率の減少を妨げることが可能になることを明らかにした(詳細は岡村他 2019 を参照)．

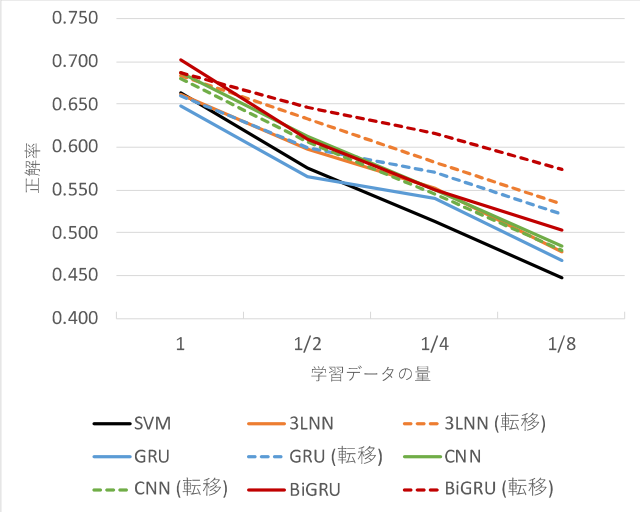

図 1 学習データを減少させた場合の意味役割付与精度の変化

意味役割付与データは人手によって作成するのはコストがかかるため学習データの量には限りがあることが予測される．そこで頻度が低い場合にも適用可能なベイズの定理に基づく学習モデルの構築を実施した．具体的には図 2 に示すように概念フレームと意味役割が出現する同時確率分布モデルを依存関係から構築した．さらに，依存関係をもとに変分下限を利用したベイズ推定を行うモデルを構築して確率分布を出力させる実験を実施した(岸本・竹内 2020 参照)．構築したモデルでは全動詞，助詞，態などの確率分布を計算するため時間が非常にかかることが明らかとなった．また概念フレームの予測では辞書の制約が無い場合には精度が低下することがあり推定が容易でないことが明らかになった．

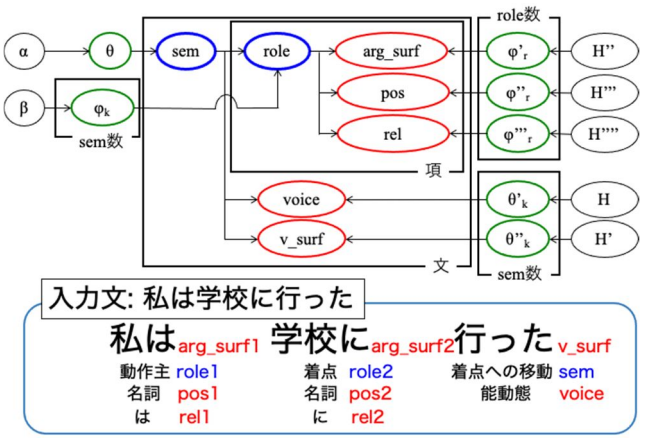

図 2 ベイズの定理を利用した概念フレームと意味役割の付与モデル

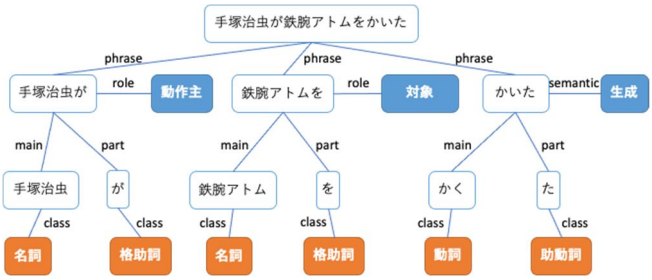

図 3 述語項構造解析後のテキストデータに対する構造化

(3) テキストに対して係り受け構造や意味役割および概念フレームが付与された際にブロックに分割されたパターンを構成して意図した検索パターンを生成する機構を構築した．具体的には

図 3 に示すように研究室で開発している述語項構造解析システム ASA (https://github.com/takenl2021/python_asa) をテキストに適用後, 文節や形態素, 品詞, 意味役割, 概念フレームをすべて木構造で記述する. 木構造はノードとリンクから構成されており, リンク部分に関係が埋め込まれている. この木構造のリンク部分を prolog の述語に置き換えることで, 文を構造化する. ユーザが文を検索する際には, 解析済みの木構造に合わせたパターンを組み合わせることで, 必要とする表現を抽出することができる (小笠原・竹内 2021 参照).

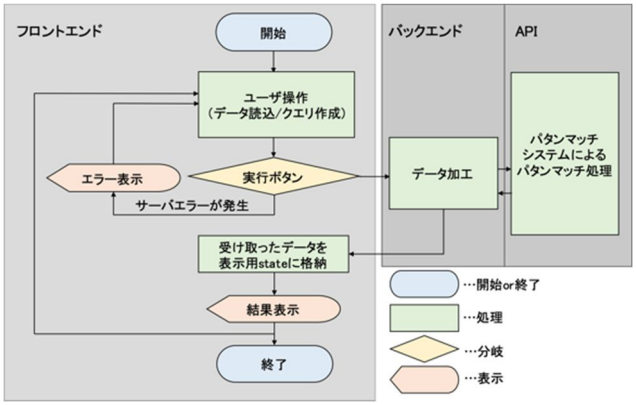

図 4 Blockly を利用したコンコーダンサの構成

(4) JavaScript を利用したコンコーダンサシステムを開発して意味役割や概念フレームを利用したパターンマッチを任意のテキストに対して実行できる環境を構築した. ブロック型のプログラミング言語として Blockly を利用して, 上述の prolog で定義された基本パターンを用意して, 組み合わせにより解析済みのテキストデータを検索することができる基本システムを開発した. システムの構成としては図 4 に示すようにユーザが操作するフロント部分と上記 (3) で記述した解析処理部分の 2 つの構成要素からなる (岡田・竹内 2021 参照).

図 5 Blockly を利用したコンコーダンサ

ユーザはテキストデータをシステムに対して upload する. テキストは解析処理部によって prolog 形式に分解される. 分解されたテキストに対して, ユーザが意味役割を利用したパターンを生成してパターンマッチを実行する. システムはマッチした文書および, マッチした部分 (項や形態素) を複数ある表示方法から選択して表示できる. 図 5 では, 著者と作品名をパターンから取り出すブロック言語を表示している (https://react.asa-gao.com/).

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 岡村 拓哉，竹内 孔一，石原 靖弘	４．巻 50
２．論文標題 ニューラルネットワークを利用した日本語意味役割付与モデルの構築	５．発行年 2019年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 2063～2074
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 オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 竹内 孔一	４．巻 69
２．論文標題 述語項構造シソーラスの構築と概念体系を利用した専門用語の処理	５．発行年 2019年
３．雑誌名 情報の科学と技術	６．最初と最後の頁 421～426
掲載論文のDOI（デジタルオブジェクト識別子） 10.18919/jkg.69.9_421	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Koichi Takeuchi, Alastair Butler, Iku Nagasaki, Takuya Okamura, Prashant Pardeshi
２．発表標題 Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus
３．学会等名 The 12th Language Resources and Evaluation Conference（国際学会）
４．発表年 2020年

１．発表者名 岸本 廉，竹内 孔一
２．発表標題 ベイズの定理を利用した日本語意味役割付与モデルの構築
３．学会等名 FIT2020(第19回情報科学技術フォーラム)
４．発表年 2020年

１．発表者名
小笠原崇，竹内孔一

２．発表標題
意味役割付与テキストに対するPrologベースの探索木による言語パタンマッチシステム構築

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
岡田魁人，竹内孔一

２．発表標題
Blocklyを利用したタグ付きコーパス検索パタン構築ツール

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
竹内孔一，アラステアバトラー，長崎郁，プラシャントパルデシ

２．発表標題
NPCMJへのPropBank形式の意味役割と概念フレームの付与の進捗報告

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
國府大輝，竹内孔一

２．発表標題
日本語WordNetにおける語義・概念の分散表現獲得

３．学会等名
FIT2019講演論文集

４．発表年
2019年

１．発表者名
Daiki Ko, Koichi Takeuchi

２．発表標題
Evaluation of Embedded Vectors for Lexemes and Synsets Toward Expansion of Japanese WordNet

３．学会等名
International Conference of the Pacific Association for Computational Linguistics（国際学会）

４．発表年
2019年

１．発表者名
竹内孔一，バトラーアラステア，長崎郁，パルデシプラシャント

２．発表標題
PropBank形式を考慮したNPCMJに対する意味役割付与～態の違いと経験者の付与～

３．学会等名
言語処理学会

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
述語項構造シソーラス
http://pth.cl.cs.okayama-u.ac.jp/
意味役割付与システム
http://www.cl.cs.okayama-u.ac.jp/study/project/asa/
コンコーダンサシステム
https://asa-gao.com/

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

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