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　 ５ 月 １２ 日現在

機関番号：１２６０５
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１４２１

研究課題名（和文）複数タグセットのタグがついたコーパスによる語義曖昧性解消の転移学習

研究課題名（英文）Transfer Learning of Word Sense Disambiguation with Corpora Tagged with Multiple Tag Sets

研究代表者

古宮　嘉那子（Komiya, Kanako）

東京農工大学・工学（系）研究科（研究院）・准教授

研究者番号：１０５９２３３９

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：複数の語義タグセットのついているコーパスを用いた語義曖昧性解消の研究を行った。
まず、二つの辞書の対応付けの研究をバイリンガル分散表現およびBERTを用いて行った。さらに、古文の語義曖昧性解消について現代文と古文の二つのタグを意識した語義曖昧性解消の研究を行った。さらに、タグの違いは単語区切りの違いと大きな関係があることから、複合語の分散表現の合成の研究をバイリンガル分散表現とニューラルネットワークのマルチタスク学習を用いて行った。さらに、それに関連して平仮名の単語分割の研究を行った。

研究成果の学術的意義や社会的意義
科研費を申請した際にはまだBERTなどの事前学習モデルは存在しなかった。そのため、複数の異なったタグセットのコーパスを利用した「語義曖昧性解消」の研究を行う予定であった。しかし、BERTの出現により翻訳などの下段タスクの前処理としての語義曖昧性解消の意義は小さくなったと考え、辞書の対応付けの研究を行うこととした。また、事前学習モデルはタガーを提供し、語彙を限定しているため、単語区切りが異なる問題があることに気づいたため、複合語の分散表現の合成の研究と平仮名の単語分割の研究を行った。さらに、古文のような言語学的観点からは、語義を知ることに意味があると考え、古文の語義曖昧性解消の研究を行った。

研究成果の概要（英文）：We conducted research on word sense disambiguation using corpora with multiple word sense tag sets.
First, we took the correspondence between two dictionaries using bilingual word embeddings and BERT. In addition, research on word sense disambiguation was conducted in historical texts with two tags, contemporary and historical tags. Furthermore, as the difference in tags sometimes come from the difference in word delimitation, we composed distributed representations of compound words from their constituent words using bilingual distributed representations and neural network multi-task learning. In addition, a related study, word segmentation in hiragana, was conducted.

研究分野：自然言語処理

キーワード：語義曖昧性解消　分散表現　対応付け　辞書　単語区切り　複合語　古文

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

## １．研究開始当初の背景

本研究課題の申請時(2017 年度)には，同一のコーパス上に，同一タスクの複数のタグセットのタグがつけられているケースが増えていた．特に，テキスト中の言葉の意味を文脈から推定するタスクである語義曖昧性解消においても，複数のタグセットの語義が付与されはじめていた．たとえば，現代日本語書き言葉均衡コーパスには岩波国語辞典の語義タグが付与された後，分類語彙表という辞書の語義（概念番号）が付与されつつあった．参照元となる辞書が異なるため，それぞれ付与すべき語義は異なるが，ともに言葉の意味を付与しているという点では一致しているため，互いに深い関連性がある．そのため，転移学習技術を用い，推定精度を高めることができると考えられた．

一方で、当時、機械学習において転移学習が注目を集めている。特に，テキスト中の人物名，組織名，地名などといった固有表現の範囲を指定し，種類を推定するタスクである固有表現抽出においては，分野によって正解率がまだ十分とはいえないものの，タグの種類および範囲が異なる場合の転移学習についての研究が報告されている (Qu et. al, EMNLP 2016;　Daniken et. al, EMNLP ワークショップ 2017)。しかし語義曖昧性解消については，このような転移学習に関する論文が報告されていなかった。理由については、

(1) 英語の研究においては語義曖昧性解消の語義は大抵 WordNet の語義を付与するものと決まっているため，転移学習をしようというモチベーションがないということと、
(2)固有表現抽出の転移学習に比べて，技術的に困難であること

が考えられた。特に後者の理由としては、コーパス中にあまり出てこない単語については，十分なサイズのコーパスがなかった。

## ２．研究の目的

語義曖昧性解消において，あるひとつのタグセット（岩波国語辞典）についての知識を，別のタグセット（分類語彙表）の学習に利用する手法を研究することを目的としていた．

具体的には以下の三手法の有効性を明らかにしようとしていた。

(a) タグの対応関係を利用する方法の有効性を明らかにする手法
(b) ソースタグセットのタグを直接素性として利用する方法
(c) 深層学習を用いた転移学習による表現ベクトルの学習の有効性

特に、あまりタグ付けがされていない語義曖昧性解消のタグセットのコーパスについて、これらの手法を用いて、その精度を高めることが最終的な目的であった。

## ３．研究の方法

上記「２．研究の目的」に記述した三手法について順に研究する予定であったが、2018 年の秋に Google 社によって BERT が発表され、2019 年の終わりか 2020 年には、大規模事前学習モデルの利用による自然言語処理が主流となってきた。BERT のような大規模事前学習モデルは大量の事例を必要とせず、少量の事例から Fine-tuning と呼ばれるもともとのモデルのパラメータの微細な調整により追加的な学習を行うことで、これまでの最高性能を大幅に上回る結果を出す。そのため、自然言語処理の研究分野の中で、タグ付きコーパスの拡張よりも、事前にタグなしのコーパスで自己学習した大規模事前学習モデルの方が有用であるという知見が得られてきた。

さらに、BERT のような Transformer 型のモデルの出力ベクトルを利用することで、語義曖昧性解消の目的とする、言葉の意味を表現したり分類したりするということがトークンごとに実現可能になり、翻訳などの下段タスの前処理としての語義曖昧性解消の重要性は下がってきたと考えている。

そのため、研究の目的自体を少々変えて、

(a) タグの対応関係を利用する方法の有効性を明らかにする手法

のほかに、

(b) 古文や平仮名の文、また日本語の科学技術論文など、大規模事前学習モデルの公開されていない分野の文書における語義曖昧性解消やその他の自然言語処理の研究を行う手法
(c) 語義タグの違いは単語区切りの違いと大きな関連があることから、複合語についての研究を行うこととした。

(a)については、Bilingual Word Embeddings を利用する手法と BERT を利用する手法について研究を行った。また、(b)についてはそれぞれのアプリケーションについて様々な手法を採用

したが、例えば、古文の語義曖昧性解消については、古典的な領域適応の手法、BERT を利用したレキシカルサンプルタスクの研究、BERT を利用した all-words WSD の手法について研究を行った。これらで利用した BERT は古文のものではなく、現代文を利用して訓練されたものである。また、LSTM を利用した場合と、T5 を利用した場合の古文からの現代日本への翻訳の研究も行った。この T5 も基本的には日本語の現代文で学習されたものを使用している。(c)については、主に、複合語の分散表現（Word2Vec）をその構成語から作成する研究を行った。この研究には、Bilingual Word Embeddings を手法に用いるものと、マルチタスク学習を用いて、言語的知識を使って、一般的な多層パーセプトロンを利用する手法をとるものについて研究を行った。

## ４．研究成果

BERT に代表されるような大規模事前学習モデルの出現は、本研究課題の申請時(2017 年度)には想定していなかったが、途中で研究の道筋の軌道修正を行うことで、「３．研究の方法」で述べた(a)～(c)の目的について、以下の知見が得られた。

(a) については、Bilingual Word Embeddings を使うことで日本語の辞書同士の語義の対応付けを教師なしで行い、2018 年に Artex らによって提唱された教師なしの対応付けの手法である VecMap の結果を上回った

(b) については、アプリケーションによってさまざまな結果が出ているが、例えば古文の語義曖昧性解消については、現代の日本語文によって訓練された BERT によるモデルが非常に有効であることが示せた。また、古文から現代文への機械翻訳について、LSTM の入力ベクトルを時代順に Fine-tuning した手法や、T5 という事前学習モデルを利用することが有効であることを示した。さらに、平仮名の単語分割については Bi-LSTM よりも疑似データを利用して作成した平仮名 BERT が有効であることを示した。

(c) については
複合語の分散表現が Bilingual Word Embeddings を使って作成できることを示した。また、複合語の分散表現がその構成語から多層パーセプトロンを用いて作成できることを示し、その学習には、品詞パターンの推定をサブタスクとしたマルチタスク学習が有効であることを示した。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　6件）

| | |
|---|---|
| １．著者名<br>Jun Izutsu, Kanako Komiya | ４．巻<br>11 |
| ２．論文標題<br>Morphological Analyzer Using the Bi-LSTM Model Only for Japanese Hiragana Sentences | ５．発行年<br>2022年 |
| ３．雑誌名<br>International Journal on Natural Language Computing | ６．最初と最後の頁<br>29-45 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.5121/ijnlc.2022.11103 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Mika Kishino, Kanako Komiya | ４．巻<br>11 |
| ２．論文標題<br>Extracting Speech Patterns of Japanese Fictional Characters Using Subword Units | ５．発行年<br>2022年 |
| ３．雑誌名<br>International Journal on Natural Language Computing | ６．最初と最後の頁<br>1-14 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.5121/ijnlc.2022.11101 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Suzuki Rui、Komiya Kanako、Asahara Masayuki、Sasaki Minoru、Shinnou Hiroyuki | ４．巻<br>26 |
| ２．論文標題<br>Unsupervised All-words WSD Using Synonyms and Embeddings | ５．発行年<br>2019年 |
| ３．雑誌名<br>Journal of Natural Language Processing | ６．最初と最後の頁<br>361～379 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.5715/jnlp.26.361 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Komiya Kanako、Kono Shinji、Seito Takumi、Hirabayashi Teruo | ４．巻<br>22 |
| ２．論文標題<br>Composing Word Embeddings for Compound Words Using Linguistic Knowledge | ５．発行年<br>2023年 |
| ３．雑誌名<br>ACM Transactions on Asian and Low-Resource Language Information Processing | ６．最初と最後の頁<br>1～22 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1145/3561299 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| 1．著者名<br>Shinji Kono、Komiya Kanako、Hiroyuki Shinnou | 4．巻<br>29 |
| 2．論文標題<br>Japanese Parsing Using Smaller BERT | 5．発行年<br>2022年 |
| 3．雑誌名<br>Journal of Natural Language Processing | 6．最初と最後の頁<br>854～874 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.5715/jnlp.29.854 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| 1．著者名<br>古宮 嘉那子、田邊 絢、新納 浩幸 | 4．巻<br>23 |
| 2．論文標題<br>分散表現を利用した日本語歴史コーパスにおける語義曖昧性解消の通時適応 | 5．発行年<br>2022年 |
| 3．雑誌名<br>国立国語研究所論集 = NINJAL Research Papers | 6．最初と最後の頁<br>59～73 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.15084/00003566 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計33件（うち招待講演　0件／うち国際学会　10件）

| |
|---|
| 1．発表者名<br>Jun Izutsu, Kanako Komiya |
| 2．発表標題<br>Morphological Analysis of Japanese Hiragana Sentences Using the Bi-LSTM CRF Model |
| 3．学会等名<br>10th International Conference on Natural Language Processing (NLP 2021)（国際学会） |
| 4．発表年<br>2021年 |

| |
|---|
| 1．発表者名<br>Mika Kishino, Kanako Komiya |
| 2．発表標題<br>Extraction of Linguistic Speech Patterns of Japanese Fictional Characters Using Subword Units |
| 3．学会等名<br>10th International Conference on Natural Language Processing (NLP 2021)（国際学会） |
| 4．発表年<br>2021年 |

１．発表者名
多喜 凪，古宮嘉那子

２．発表標題
現代文 BERT を利用した日本語歴史コーパスの語義曖昧性解消

３．学会等名
言語処理学会第28回年次大会

４．発表年
2021年

１．発表者名
河野稜斗，平林照雄，古宮嘉那子

２．発表標題
BERTを用いた二つの辞書の対応付け

３．学会等名
言語処理学会第28回年次大会

４．発表年
2021年

１．発表者名
三戸尚樹，古宮嘉那子，佐々木稔

２．発表標題
共学習によるレビュー文書からのネガティブな意見文の抽出

３．学会等名
言語処理学会第28回年次大会

４．発表年
2021年

１．発表者名
HUANG YIPU，佐々木稔，古宮嘉那子

２．発表標題
レビューから抽出されたキーフレーズと感情スコアを用いた評判分析

３．学会等名
言語処理学会第28回年次大会

４．発表年
2021年

１．発表者名
Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara and Hiroyuki Shinnou

２．発表標題
Automatic Creation of Correspondence Table of Meaning Tags from Two Dictionaries in One Language Using Bilingual Word Embedding

３．学会等名
13th BUCC Workshop at LREC 2020（国際学会）

４．発表年
2020年

１．発表者名
Kanako Komiya, Daiki Yaginuma, Masayuki Asahara, Hiroyuki Shinnou

２．発表標題
Generation and Evaluation of Concept Embeddings Via Fine-Tuning Using Automatically Tagged Corpus

３．学会等名
PACLIC 2020（国際学会）

４．発表年
2020年

１．発表者名
Teruo Hirabayashi, Kanako Komiya, Masayuki Asahara

２．発表標題
Composing Word Vectors for Japanese Compound Words Using Dependency Relations

３．学会等名
PACLIC 2020（国際学会）

４．発表年
2020年

１．発表者名
Masashi Takaku, Tosho Hirasawa, Mamoru Komachi, Kanako Komiya

２．発表標題
Neural Machine Translation from Historical Japanese to Contemporary Japanese Using Diachronically Domain-Adapted Word Embeddings

３．学会等名
PACLIC 2020（国際学会）

４．発表年
2020年

１．発表者名
佐々木稔，古宮嘉那子

２．発表標題
複数の事前学習済みモデルを用いたQAサイト質問回答ペアの分類

３．学会等名
IDRユーザフォーラム 2020

４．発表年
2020年

１．発表者名
河野 慎司，古宮嘉那子

２．発表標題
品詞情報を利用した複合語の分散表現の合成

３．学会等名
音声言語および自然言語処理シンポジウム

４．発表年
2020年

１．発表者名
井筒順，古宮嘉那子

２．発表標題
Bi-LSTM CRF モデルを用いた平仮名文の形態素解析

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
平林照雄，河野慎司，古宮嘉那子，新納浩幸

２．発表標題
日本語の論文コーパスにおける「問題」の語義アノテーション

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
岸野望叶，古宮嘉那子

２．発表標題
SentencePieceを用いたキャラクターの特徴語抽出

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
金野佑太，古宮嘉那子

２．発表標題
論文の要旨からのタイトル生成におけるキーワードと分野別fine-tuningの効果

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
小林汰一郎，古宮嘉那子

２．発表標題
SVMを用いたBCCWJにおける同形異音語の読み推定

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
Kanako Komiya, Takumi Seitou, Minoru Sasaki, Hiroyuki Shinnou

２．発表標題
Composing Word Vectors for Japanese Compound Words Using Dependency Relations

３．学会等名
CICLING 2019（国際学会）

４．発表年
2019年

１．発表者名
柳沼 大輝，古宮 嘉那子，新納 浩幸

２．発表標題
All-words WSDとfine-tuningを利用した分類語彙表の語義の分散表現の構築

３．学会等名
言語資源活用ワークショップ 2019

４．発表年
2019年

１．発表者名
佐々木稔， 古宮嘉那子

２．発表標題
単語区切りの違いによるQAサイトの質問回答ペアの分類

３．学会等名
IDRユーザフォーラム 2019

４．発表年
2019年

１．発表者名
井筒順，明石陸，加藤涼，岸野望叶，小林汰一郎，金野佑太，古宮嘉那子

２．発表標題
MeCab による平仮名のみの形態素解析

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
河野慎司，古宮嘉那子

２．発表標題
マルチタスク学習を利用した短単位の分散表現から長単位の分散表現の合成

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
高久雅史，平澤寅庄，小町守 ，古宮嘉那子

２．発表標題
通時的な領域適応を行った単語分散表現を利用した古文から現代文へのニューラル機械翻訳

３．学会等名
言語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
平林照雄，古宮嘉那子，新納浩幸

２．発表標題
Bilingual Word Embeddingsによる短単位と長単位のアラインメント

３．学会等名
語処理学会第26回年次大会

４．発表年
2020年

１．発表者名
Masaya Suzuki, Kanako Komiya, Minoru Sasaki and Hiroyuki Shinnou

２．発表標題
Fine-tuning for Named Entity Recognition Using Part-of-Speech Tagging

３．学会等名
The 32nd Pacific Asia Conference on  Language, Information and Computation（国際学会）

４．発表年
2018年

１．発表者名
Aya Tanabe, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou

２．発表標題
Detecting Unknown Word Senses in Contemporary Japanese Dictionary from Corpus of Historical Japanese

３．学会等名
Japanese Association for Digital Humanities 2018（国際学会）

４．発表年
2018年

１．発表者名
平林照雄，古宮 嘉那子，新納浩幸

２．発表標題
Bilingual Word Embeddingsによる『岩波国語辞典』の語義と『分類語彙表』の語義の対応付け

３．学会等名
言語処理学会第25回年次大会

４．発表年
2019年

１．発表者名
Kanako Komiya, Nagi Oki and Masayuki Asahara

２．発表標題
Word Sense Disambiguation of Corpus of Historical Japanese Using Japanese BERT Trained with Contemporary Texts

３．学会等名
The 36th Pacific Asia Conference on Language, Information and Computation（国際学会）

４．発表年
2022年

１．発表者名
臼井久生，古宮嘉那子

２．発表標題
T5を用いた古文から現代文への翻訳

３．学会等名
言語処理学会第29回年次大会

４．発表年
2022年

１．発表者名
浅田宗磨，古宮嘉那子

２．発表標題
日本語歴史コーパスのAll-words WSD

３．学会等名
言語処理学会第29回年次大会

４．発表年
2022年

１．発表者名
井筒順， 古宮嘉那子，新納浩幸

２．発表標題
平仮名BERTを用いた平仮名文の分割

３．学会等名
言語処理学会第29回年次大会

４．発表年
2022年

１．発表者名
井筒順，古宮嘉那子， 新納浩幸

２．発表標題
平仮名BERTによる平仮名文の分割

３．学会等名
第253回自然言語処理研究発表会

４．発表年
2022年

１．発表者名
平林照雄，古宮嘉那子,浅原正幸

２．発表標題
科学技術論文における「問題」の周辺文からの問題内容の抽出

３．学会等名
言語資源ワークショップ2022

４．発表年
2022年

〔図書〕　計1件

| １．著者名 | ４．発行年 |
|---|---|
| 柴原 一友、築地 毅、古宮 嘉那子、宮武孝尚、小谷 善行 | 2019年 |
| ２．出版社 | ５．総ページ数 |
| 森北出版 | 240 |
| ３．書名 | |
| 機械学習教本 | |

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（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 |
|---|---|