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　 ６ 月 ２３ 日現在

機関番号：１２１０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１４２２

研究課題名（和文）半教師あり深層学習を用いた語義曖昧性解消

研究課題名（英文）Word Sense Disambiguation Using Semi-supervised Deep Learning

研究代表者
佐々木　稔（Sasaki, Minoru）

茨城大学・理工学研究科（工学野）・准教授

研究者番号：６０３４４８３４

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究は半教師ありディープラーニングを用いて対象単語前後の単語からなる特徴ベクトルと用例文間の関係を表すグラフ埋め込みベクトルによる高精度な語義曖昧性解消システムの開発を行った。システムの有効性を評価した結果、開発したシステムは既存の日本語半教師あり語義曖昧性解消システムと比較して、語義識別の精度が1.73%向上した。また、英語の評価データであるSENSEVAL-2 English Lexical Taskデータを使用して語義曖昧性解消実験を行った結果、最高精度が得られた従来手法と比較して精度が3%向上した。これらの結果より開発システムが語義曖昧性解消に有効であることを示すことができた。

研究成果の学術的意義や社会的意義
語義曖昧性解消において、「語義曖昧性解消をシンプルな半教師ありディープラーニングを使ったモデルで構築できないか」「少量の語義付き用例文を利用して語義の特徴を捉えたディープラーニングモデルを構築できないか」という2つの課題を解決する効果的な手法を確立することができた。
本研究の成果から得られる学術的な意義は、語義付き例文が少量のみ存在する場合でも従来手法では捉えられなかった効果的な文脈情報の取得や用例文間の意味的な関係の取得が可能となったことである。この成果により、用例文を大量に追加して効果的な識別モデルの学習が可能なことや用例文を大量に提供可能な国語辞典の編纂が可能となるなどの社会的意義がある。

研究成果の概要（英文）：In this study, we developed a semi-supervised WSD method using semantic similarities between example sentences. In this method, we propose a graph construction method that does not require any parameters using BERT pre-trained model to represent a semantic similarity relation obtained from sense labeled examples and unlabeled examples. As a result of evaluating the effectiveness of the system, the developed system improved the accuracy of word sense identification by 1.73% compared to an existing Japanese semi-supervised word sense disambiguation system. In addition, the results of a word sense disambiguation experiment using the SENSEVAL-2 English Lexical Task data, which is English assessment data, showed a 3% improvement in accuracy compared to the previous method, which achieved the highest accuracy. These results show that the developed system is effective in semi-supervised word sense disambiguation.

研究分野：自然言語処理

キーワード：語義曖昧性解消　機械学習　グラフニューラルネットワーク　半教師あり学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

一般的な自然言語処理システムは単語を低次元の数値ベクトルで表現し、特徴を捉えている。このベクトルはニューラルネットワークなどの機械学習システムの入力として使用され、語義曖昧性解消などの様々なタスクを解くための手掛かりとなっている。近年、機械学習に基づく語義曖昧性解消はディープラーニングを用いる手法が主流となり、現在では多義語をより効果的に識別するための特徴抽出手法の開発が主要な研究課題となっている。

ディープラーニングに基づく語義曖昧性解消は、2013 年に Mikolov が文書中の単語をベクトルで表現する word2vec を開発したことがきっかけとなり、単語ベクトルが様々な識別モデルへの入力に取り入れられた。例えば対象の多義語に対し、辞書の語義説明文から求めた語義ベクトルと周辺で共起する単語から求めた文脈ベクトルを比較し、語義曖昧性解消を行う教師なし手法(Chen et al. EMNLP2014)や対象単語を中心とした前後 N 単語の単語ベクトルを連結した文脈ベクトルを用いて教師あり学習を行う手法(Sugawara et at. PACLING2015)などの多数の手法が存在する。半教師あり学習手法では単語の予測モデルへの入力として、ラベル伝搬法で語義を推定した用例文を入力し、各語義が取りやすい単語を予測する手法が提案されている(Yuan et al. COLING2016)。

教師あり学習手法は高い精度で分類することができるが、訓練データとなる語義付き用例文が大量に必要となる。ディープラーニングで学習をするには少量のデータでは性能が高くないことが知られている。しかし、語義を付与するには文脈を理解する必要があるため専門的知識が求められ、大量のデータを人手で付与することから、訓練データの作成は時間がかかり、大量に用意することは困難である。一方で、教師なし学習手法は大量に存在する語義なし用例文集合を使い、用例文のパターンや語義別カテゴリへの自動分類が可能である。大量の用例文集合があれば語義の用法や特徴を発見するため、語義付き用例文を用意する必要はない。しかし、教師なし学習手法は出力された語義カテゴリがどのような基準で分類されたかについて理解ができない場合がある。また、分類するカテゴリ数を事前に指定することが難しいという欠点も存在する。このような理由から、少量の語義付き用例と大量の語義なし用例を用いて語義の識別精度の高いモデルを構築できる、半教師ありのディープラーニングに基づく語義曖昧性解消を行うことに注目した。

ディープラーニングに基づく半教師あり学習を用いた語義曖昧性解消手法はこれまでに 2 件の手法しか報告されていない。その手法は教師なし学習と教師あり学習の組合せ(Taghipour et al. NAACL2015)、教師なし、半教師あり、教師あり学習の組合せを用いたモデル(Yuan et al. COLING2016)であるが、非常に複雑なモデルとなっている。また、ラベルとして使用する辞書の意味区分に合わせたニューラルネットワークの学習ができていないことも問題である。従来手法は教師なし学習を用いているため、頻繁に出現する用例文のパターンを捉えることはできるが、使用する辞書の語義に対応したパターンが得られるとは限らない。以上のように、「語義曖昧性解消をシンプルな半教師ありディープラーニングを使ったモデルで構築できないか」と「少量の語義付き用例文を利用して語義の特徴を捉えたディープラーニングモデルを構築できないか」という 2 つの問いに対して、語義の意味区分を考慮したシンプルな半教師ありディープラーニングを用いた語義曖昧性解消システムを構築する

ことが本研究の中心的な課題である。

2．研究の目的

本研究の目的は以下の 3 点である。

（1） **高精度な半教師ありディープラーニングを用いた語義曖昧性解消システムの開発**

語義曖昧性解消と用例文間関係の学習を同時に行うシステムはこれまでに存在せず、新しいアプローチで語義曖昧性解消を行う。深層学習によって語義付き用例文と語義なし用例文の有効な特徴を捉える点も、これまでにないアイデアである。

（2） **シソーラスや係り受け情報を用いた語義なし用例文間の意味的な関係の調整**

ラベル伝搬法による半教師あり学習では、類似度行列からグラフ構造を求めるのが一般的である。本研究ではシソーラスや係り受け情報などの情報を事前に与えることができる点で独自性がある。

（3） **半教師あり学習結果の自己学習による語義曖昧性解消の精度向上**

(１)と(２)でシステムが出力した用例文の語義を自動的にディープラーニングにフィードバックさせることでモデルの改善を行うのは独自のアイデアである。訓練データに少量しかない語義について用例文を追加することで、バランスの取れた語義識別モデルを構築できることも特色である。

3．研究の方法

本研究では以下の 3 点について研究を行った。

(１) **高精度な半教師ありディープラーニングを用いた語義曖昧性解消システムの開発**

高精度の語義曖昧性解消システムを構築するために、語義の識別と用例文間の関係を同時に学習する半教師あり学習手法を開発する(図1)。この手法を用いて、語義曖昧性解消の精度が既存手法よりも向上することを明らかにする。この手法は対象単語の語義付き用例文集合と語義なし用例文集合の他に、用例文間の関係(類似度行列)を表すグラフ構造を入力とする。用例文に対して対象単語の周辺に共起する単語の頻度を計算し、文脈ベクトルに変換する。すべての文脈ベクトルを用いてラベル伝搬法に類似する方法で用例文間の関係を各隠れ層で学習し、その関係と文脈ベクトルを用いて予測語義の学習を行う。テストデータも同様に学習されたモデルに入力することで、対象単語の語義を予測することができる。システムの有効性を幅広く評価するため、本研究は日本語と英語の評価セットを用いて実験を行う。日本語はSemeval2010日本語タスクのデータ、英語では数多くの論文で使われるSemeval2007 lexical sample task データを使用し、すべての評価セットで精度が向上することを示す。

（2） **シソーラスや係り受け情報を用いた語義なし用例文間の意味的な関係の調整**

(１)で得られた学習モデルにおけるグラフ学習用の隠れ層に対して、シソーラスや係り受け情報などの外部情報を用いて語義識別の有効性を高めるように学習モデルの改良を行う。この改良によって、(１)で構築した語義曖昧性解消システムの識別精度が向上し、シソーラスや係り受け情報などの外部情報が有用であることを明らかにする。シソーラスから得られる単語間の上位下位関係や類義語関係、係り受け情報から得られる詳細な共起関係をグラフ構造に反映することで、有効性をさらに向上した用例文間の意味的な関係を構築する。また、用例文間の関係ではなく、単語の共起グラフを

用いて外部情報を反映させるアプローチについても有用性の検討を行う。(１)で構築した語義曖昧性解消システムの精度向上に対して、類義語関係や係り受け情報を利用しやすい共起グラフが有用なグラフ構造として利用可能であることを示す。

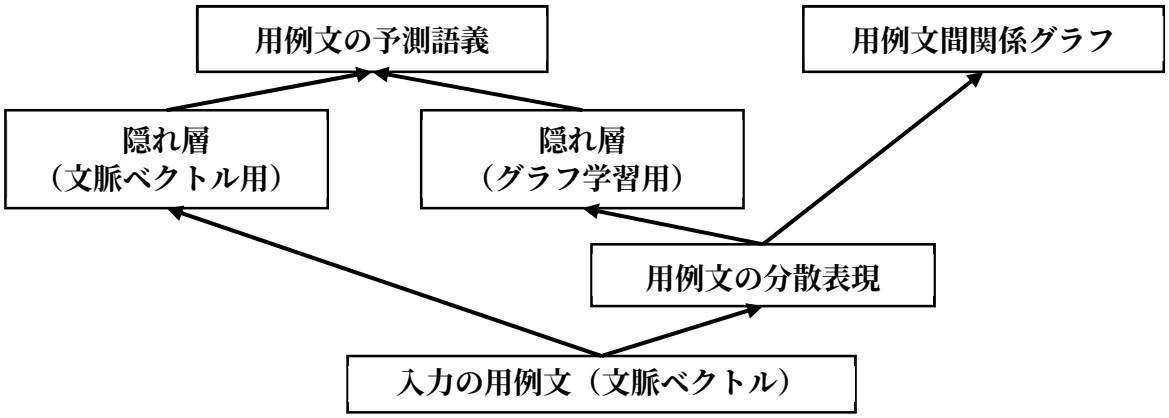

図１：語義識別と用例文間関係の学習を行う半教師あり学習モデル

（３） 半教師あり学習結果の自己学習による語義曖昧性解消の精度向上

（１)と(２)でシステムが出力した用例文の語義を自動的にディープラーニングにフィードバックさせることで、各語義に数件ずつの語義付き用例文を含む訓練データで語義曖昧性解消を高い精度で識別できることを明らかにする。(２)では外部情報を用いてグラフの特徴を表現する隠れ層の改善を行ったが、外部情報では文脈ベクトルの特徴を捉える隠れ層の改善は行われない。文脈ベクトルの隠れ層を学習するためには大量の語義付き用例文が必要となるが、コストがかかるため少量だけで高い精度の識別が可能となるシステムが必要となる。そこで半教師あり学習で得られた結果を訓練データとして利用することで、語義曖昧性解消に効果がある文脈的な特徴を抽出する。

４．研究成果

半教師ありディープラーニングに基づく語義曖昧性解消は大量の文書集合における単語の使用法と辞書の語義情報の両方を特徴として捉えるモデルとして期待されるが、これまでに２件の手法しか存在していない。しかし、これらの手法には「語義曖昧性解消をシンプルな半教師ありディープラーニングを使ったモデルで構築できないか」「少量の語義付き用例文を利用して語義の特徴を捉えたディープラーニングモデルを構築できないか」という未解決の課題が存在する。そこで、本研究は半教師ありディープラーニングを用いて対象単語前後の単語からなる特徴ベクトルと用例文間の関係を表すグラフ埋め込みベクトルによる高精度な語義曖昧性解消システムの開発を行った。入力した用例文に対して学習済み言語モデルを用いて得られた特徴ベクトルと、用例文間の関係(類似度行列)を表すグラフ構造に対してグラフ埋め込みを学習して得られたグラフのベクトルを計算し、これらを連結したベクトルから適切な語義を出力できるニューラルネットワークの学習を行う。テストデータも同様に学習されたモデルに入力することで、対象単語の語義を予測することができる。

開発した半教師あり語義曖昧性解消システムの有効性を評価するため、日本語の評価データである Semeval2010 日本語タスクデータを使用し、語義曖昧性解消実験を行っ

た。その結果、開発したシステムは既存の日本語半教師あり語義曖昧性解消システムと比較して、語義識別の精度が 1.73%向上した。また、英語の評価データである SENSEVAL-2 English Lexical Task データを使用して語義曖昧性解消実験を行った結果、最高精度が得られた従来手法と比較して精度が 3%向上した。これらの結果より開発システムが語義曖昧性解消に有効であることを示すことができた。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3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 谷田部梨恵，佐々木稔	４．巻 62
２．論文標題 用例文間の意味的な類似関係を用いた半教師あり語義曖昧性解消	５．発行年 2021年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 1724～1736
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 Suzuki Rui、Komiya Kanako、Asahara Masayuki、Sasaki Minoru、Shinnou Hiroyuki	４．巻 26
２．論文標題 Unsupervised All-words WSD Using Synonyms and Embeddings	５．発行年 2019年
３．雑誌名 Journal of Natural Language Processing	６．最初と最後の頁 361～379
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.26.361	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 Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, Hiroyuki Shinnou	４．巻 34
２．論文標題 Comparison of Methods to Annotate Named Entity Corpora	５．発行年 2018年
３．雑誌名 Transactions on Asian and Low-Resource Language Information Processing	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3218820	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 －

〔学会発表〕　計22件（うち招待講演　1件／うち国際学会　11件）

１．発表者名 Hiroki Okemoto, Minoru Sasaki
２．発表標題 Japanese Word Sense Disambiguation Using Gloss Information of a Japanese Dictionary
３．学会等名 the Thirteenth International Conference on Information, Process, and Knowledge Management (eKnow2021)（国際学会）
４．発表年 2021年～2022年

１．発表者名
Kazuki Oda, Minoru Sasaki

２．発表標題
Person Name Extraction from TV program Using Pre-trained Language Model and News Headlines

３．学会等名
the 12th International Conference on E-Service and Knowledge Management (ESKM 2021)（国際学会）

４．発表年
2021年～2022年

１．発表者名
Minoru Sasaki

２．発表標題
The reliability of word meanings in online dictionaries and how word meanings change over time

３．学会等名
The Thirteenth International Conference on Pervasive Patterns and Applications (PATTERNS2021)（招待講演）（国際学会）

４．発表年
2021年～2022年

１．発表者名
石井佑樹，佐々木稔

２．発表標題
辞書の階層構造埋め込み学習における日本語辞書定義文の効果的な利用

３．学会等名
言語処理学会第26回年次大会

４．発表年
2021年～2022年

１．発表者名
関谷洸，佐々木稔

２．発表標題
語義の例文を使用した語義曖昧性解消の有効性分析

３．学会等名
言語処理学会第26回年次大会

４．発表年
2021年

１．発表者名
谷田部梨恵，佐々木稔

２．発表標題
訓練事例と辞書用例を異なるモデルで表現した語義曖昧性解消

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
Rie Yatabe, Minoru Sasaki

２．発表標題
Semi-supervised Word Sense Disambiguation Using Example Similarity Graph

３．学会等名
Proceedings of the 14th Workshop on Graph-Based Natural Language Processing (TextGraphs-14)（国際学会）

４．発表年
2020年

１．発表者名
Rie Yatabe, Minoru Sasaki

２．発表標題
Word Sense Disambiguation Using Graph-based Semi-supervised Learning

３．学会等名
Proceedings of The Fourteenth International Conference on Advances in Semantic Processing (SEMAPRO2020)

４．発表年
2020年

１．発表者名
佐々木稔，谷田部梨恵

２．発表標題
語義曖昧性解消における辞書に定義された単義語利用についての分析

３．学会等名
言語資源活用ワークショップ2020

４．発表年
2020年

１．発表者名
Minoru Sasaki

２．発表標題
Active Learning to Select Unlabeled Examples with Effective Features for Document Classification

３．学会等名
The 10th International Conference on Computational Linguistics and Intelligent Text Processing（国際学会）

４．発表年
2019年

１．発表者名
Minoru Sasaki, Tetsuya Nogami

２．発表標題
Ibrk at the NTCIR-14 QA Lab-PoliInfo Classification Task

３．学会等名
The Fourteenth NTCIR conference (NTCIR-14)（国際学会）

４．発表年
2019年

１．発表者名
史文愷，細木唯以，三好勝博，江口潤一，佐々木稔，鈴木智也

２．発表標題
BERTモデルとニュースヘッドラインによるAI運用システムの試作

３．学会等名
日本機械学会2019年茨城講演会

４．発表年
2019年

１．発表者名
谷田部梨恵，佐々木稔

２．発表標題
グラフニューラルネットワークを用いた半教師あり語義曖昧性解消

３．学会等名
情報処理学会 第241回自然言語処理研究会

４．発表年
2019年

１．発表者名
谷田部梨恵，佐々木稔

２．発表標題
半教師あり語義曖昧性解消における各ジャンルの語義なし用例文の利用

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
谷田部梨恵，佐々木稔

２．発表標題
BERTの学習済みモデルを用いた用例文ペアの同義判定

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
木村泰知，渋木英潔，高丸圭一 ，秋葉友良，石下円香，内田ゆず，小川泰弘，乙武北斗，佐々木稔，三田村照子，横手健一，吉岡真治，神門典子

２．発表標題
NTCIR-15 QA Lab-PoliInfo2 のタスク設計

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
Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, Hiroyuki Shinnou

２．発表標題
All-words Word Sense Disambiguation Using Concept Embeddings

３．学会等名
Proceedings of the 11th edition of the Language Resources and Evaluation Conference（国際学会）

４．発表年
2018年

１．発表者名
Aya Tanabe, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, 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
The 8th Conference of Japanese Association for Digital Humanities（国際学会）

４．発表年
2018年

１．発表者名
Minoru Sasaki

２．発表標題
Multi-Domain Word Embeddings for Semantic Relation Analysis among Domains

３．学会等名
Proceedings of The Fourth Asia Pacific Corpus Linguistics Conference（国際学会）

４．発表年
2018年

１．発表者名
Minoru Sasaki

２．発表標題
Word Embeddings of Monosemous Words in Dictionary for Word Sense Disambiguation

３．学会等名
Proceedings of The Twelfth International Conference on Advances in Semantic Processing（国際学会）

４．発表年
2018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 32th Pacific Asia Conference on Language, Information and Computation（国際学会）
４．発表年 2018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

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