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 １１ 日現在

機関番号：１２１０１
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０４６

研究課題名（和文）局所的な周辺文脈を利用した日本語の教師なしAll-words型語義曖昧性解消

研究課題名（英文）All-words Word Sense Disambiguation in Japanese using Local Context

研究代表者
古宮　嘉那子（Komiya, Kanako）

茨城大学・理工学研究科（工学野）・講師

研究者番号：１０５９２３３９

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）：日本語の文章を対象に，文中の全ての単語が辞書のどの意味を持っているのかを同定する仕組みを作成した．この仕組みは教師なし学習といって，用例文とその正解を用いず，ただの平文の文例集（コーパス）だけで実行が可能である．
具体的には，意味を知りたい単語の周りの単語の意味ベクトル（ディープ･ラーニングによって算出したもの）と，意味を知りたい単語の類義語を利用することで，実現した．

研究成果の学術的意義や社会的意義
日本語を対象としたall-wordsの語義曖昧性解消（文中の全ての単語を対象として単語が辞書のどの意味を持っているのかを同定するタスク）を行う仕組みを提案した．文例集と正解を与えなくても語義曖昧性解消を行う仕組みを提案することができた．
また，この際に使うディープ･ラーニング技術の適切なパラメータや繰り返し回数についても実験によって明らかにした．

研究成果の概要（英文）：We developed an unsupervised algorithm to identify the meaning of every word in a corpus. The target language is Japanese.
We used surrounding word vectors, which are the word embeddings generated via deep learning method and synonyms of the target words.

研究分野：自然言語処理

キーワード：語義曖昧性解消　教師なし学習　All-words

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

語義曖昧性解消の手法は，大きく教師あり学習と教師なし学習の二つに分けられる。教師有り学習では高い精度で多義語の語義を推定することが可能だが，学習データの作成には高い人的コストがかかるため，あらゆる多義語に対応できる量のデータを用意することは不可能である．そのため，文中の語義タグが付与された全ての単語を同時に対象にして語義曖昧性解消を行う，All-words 型の語義曖昧性解消では，教師無しの手法をとるのが一般的である．
英語を対象にした All-words 型の教師無し語義曖昧性解消に関する研究は多く，様々な手法が考えられているが，中でもトピックモデルを用いた手法は有力である (JBoyd-Graber et al.，EMNLP2007)，(Guo et al.，EMNLP 2011)．これに対し、対象語を日本語に限ると、教師なしの All-words 型語義曖昧性解消はほとんど行われておらず（(Baldwin et. al.,IJCNLP 2008)のみ），中でもトピックモデルのような確率的に文脈を利用したものは存在しなかった．

２．研究の目的

本研究は、日本語の教師なし語義曖昧性解消の研究である。語義タグのついていないコーパスと、オントロジー型の概念辞書を用いて、周辺文脈の語義から、全単語の語義曖昧性解消を同時に行う．本研究の具体的な目的は，１）これまで行われてこなかった、日本語の All-words 型の教師なし語義曖昧性解消を行うことと，２）そのために、局所的な周辺文脈の語義を利用したモデルを構築・発展させることの二つであった．

３．研究の方法

平成 27 年は，事前実験として行っていた手法の結果をまとめ，国際会議で発表を行い，同時に国際会議において，当時の研究動向について調査を行った．事前実験として行っていた手法は，各語義が持つ周辺語義に関する確率分布を概念辞書における WORDNET-WALK（概念階層上を確率に従って遷移することにより，語義を生成するモデル）に置き換えて，各語義の確率をギブスサンプリングにより計算し、語義を決定する「周辺語義モデル」である（Komiya et al, 2015）．このシステムは，EDR コーパスを利用したものであった．深層学習の研究が増えてきつつあり，また適切に利用すれば本研究にも役立つと思われたため，今後はトピックモデルよりも深層学習を使うこととした．

平成 28 年は，自然言語処理のトップカンファレンスである，ACL2016 に参加して，最先端の自然言語処理の研究について知識を深めた．それと同時に，国立国語研究所から提供いただいた，分類語意表の意味タグのついた書き BCCWJ（現代日本語書き言葉均衡コーパス）を利用して研究を行うことにした．ギブスサンプリングを使った周辺語義の確率の代わりに，深層学習を利用した技術である分散表現を利用して，周辺の文脈の語義を表現し，類義語の情報から，対象単語の語義を同定する手法を提案した．コーパスのアノテーション（タグ付け）作業が途中であったため，この際に利用したコーパスはまだ少量であった．そのため，分散表現を計算するためのコーパスには，別コーパスを利用していた．当初は，タグなしの BCCWJ を利用していたが，国立国語研究所の浅原准教授（当時）に大量の Web 文書からなる分散表現を作成していただいたため，後にそちらに置き換えた．このときの成果が（鈴木 et al，2017.03）である．

平成 29 年度は，パラメータ等をさまざまに試しつつ，本手法を練り上げて，きれいな形で実験を行って，考察を深くした．また，その結果を平成 29 年度 11 月に国際会議に投稿した．この成果が，（Suzuki et al, 2018）である．また，前年度に副産物としてできた Web 文書からなる分散表現関連について議論を行い知見を深めた．

平成 30 年度は，投稿した論文の発表を行った後，アノテーション作業の終わった BCCWJ を使って大規模な実験を行い，その結果をまとめてジャーナル論文に投稿した．この成果が（鈴木 et al，2019）である．その結果，「自然言語処理」令和元年度の六月号に掲載が決まった．

４．研究成果

５節に示すとおり，研究目的に沿った研究およびその関連研究として，雑誌論文を６件と，学会発表を５０件行った．特に，「周辺語義モデル」による all-words の語義曖昧性解消に加え，分散表現と類義語を利用した all-words の語義曖昧性解消についての手法を提案した．

ひとつめの「周辺語義モデル」は，各語義が持つ周辺語義に関する確率分布を概念辞書における WORDNET-WALK に置き換えて，各語義の確率をギブスサンプリングにより計算し、語義を決定するモデルであり，国際会議で発表した（Komiya et al, 2015）．

ふたつめの分散表現と類義語を利用した all-words の語義曖昧性解消は，深層学習を用いた基幹技術である分散表現を単語ごとに計算し，さらに語義曖昧性解消の対象単語と，その類義語の周辺の単語の意味的な近さを分散表現で計算することによって語義曖昧性解消を行うモデルである．この手法に関しては，（鈴木 et al, 2017）で研究会の発表を行い，（Suzuki et al, 2018）でより詳細な実験を行って国際会議で発表し，（鈴木 et al. 2019）でさらに新しいコーパスを使って実験した後，より詳細に考察を加えて雑誌論文として発表した．

これらの手法によって，日本語のテキストから，教師データ（語義タグデータ）を一切用いずに全ての単語の語義を予測する，all-words の語義曖昧性解消が行えるようになった.

５．主な発表論文等

〔雑誌論文〕（計　6　件）

（１） 鈴木類，古宮嘉那子，浅原正幸，佐々木稔，新納浩幸，概念辞書の類義語と分散表現を利用した教師なし all-words WSD，自然言語処理，Vol.26，No.2，(to appear)，(2019.6)．（査読有）

（２） Kanako Komiya, Minoru Sasaki, Hiroyuki Shinnou, Manabu Okumura, Domain Adaptation using Word Embeddings for Word Sense Disambiguation，自然言語処理，Vol.25，No.4 pp.463-480,(2018.9)．　（査読有）

（３） Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, Hiroyuki Shinnou, Comparison of Methods to Annotate Named Entity Corpora, Transactions on Asian and Low-Resource Language Information Processing, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 17 Issue 4, Article No. 34,(2018.8).（査読有）

（４） Kanako Komiya, Minoru Sasaki, Hiroyuki Shinnou, Yoshiyuki Kotani, Cross-lingual Product Recommendation System Using Collaborative Filtering，自然言語処理，Vol. 24，No. 4，pp. 579-596，(2017.9).（査読有）

〔学会発表〕（計 50　　件）

（１） Kanako Komiya, Hiroyuki Shinnou, Investigating Effective Parameters for Fine-tuning of Word Embeddings Using Only a Small Corpus, DeepLo 2018, Workshop of ACL 2018, (2018.7).（査読有）

（２） Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou, All-words Word Sense Disambiguation Using Concept Embeddings, LREC 2018, (2018.5).（査読有）

（３） Kanako Komiya, Minoru Sasaki, and Hiroyuki Shinnou, Comparison of Distributed Representations and Context Vectors for Japanese Onomatopoeia Classification, CICLING 2018, (2018.3).（査読有）

（４） Kanako Komiya, Shota Suzuki, Minoru Sasaki, Hiroyuki Shinnou, and Manabu Okumura, Domain Adaptation for Word Sense Disambiguation Using Word Embeddings, CICLING 2017, LNCS 10762, Computational Linguistics and Intelligent Text Processing, (2017.4).（査読有）

（５） 鈴木類，古宮 嘉那子，浅原 正幸，佐々木 稔，新納 浩幸，『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消，言語処理学会第 23 回年次大会，(2017,03,14).（査読なし）

（６） Kanako Komiya, Minoru Sasaki, Hiroyuki Shinnou, Yoshiyuki Kotani, and Manabu Okumura, Selecting Training Data for Unsupervised Domain Adaptation in Word Sense Disambiguation, The Pacific Rim International Conference on Artificial Intelligence (PRICAI) 2016: Trends in Artificial Intelligence, Lecture Notes in Computer Science LNAI9810, (2016.8).（査読有）

（７） Kanako Komiya, Masaya SUZUKI, Tomoya Iwakura, Minoru Sasaki, Hiroyuki Shinnou, Comparison of Annotating Methods for Named Entity Corpora, LAW-X 2016, Workshop of ACL 2016, (2016.8).（査読有）

（８） Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani, Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation, PACLIC 2015, (2015.10).

〔図書〕（計　0　件）

〔産業財産権〕
なし

〔その他〕
なし

６．研究組織

(1)研究分担者
なし

(2)研究協力者
研究分担者氏名：鈴木類
ローマ字氏名：Rui Suzuki
所属研究機関名：茨城大学
部局名：工学部
職名：学生
研究者番号（8 桁）：なし