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　 ５ 月 １３ 日現在

機関番号：１３５０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2021
課題番号：１７Ｋ００２９９

研究課題名（和文）語義の大域的特徴量とヒトの修正過程に基づく時系列文書の自動分類

研究課題名（英文）Learning Timeline Difference for Text Categorization based on Global Features of Word Senses and Category Modification

研究代表者

福本　文代（Fukumoto, Fumiyo）

山梨大学・大学院総合研究部・教授

研究者番号：６０２６２６４８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究は，訓練文書と作成時期が異なるテスト文書を分類するために有効な語彙的意味処理技術と教師付き学習手法を開発することを目的とする．今年度はこれまで実施してきた研究の成果を論文として5編公開した．具体的には，(1) 分野語義に関する論文として2編，(2) 語義解消に関して提案した2手法をそれぞれ1編，そして (3) 意味の等価性に関する論文である．さらに，文書に付与された分野語義と文書中との意味の同値性を用いて文書を高精度で分類する手法について提案した．今後は少量の訓練データについても高精度な分類が行えるよう，分野依存語義適応に関する手法を検討する予定である．

研究成果の学術的意義や社会的意義
本研究で提案する分野語義獲得手法は，人手で作成した辞書とは異なり辞書の改版や異なる体系を持つ辞書，異なる言語辞書に対しても柔軟に適用可能な枠組みであり，このことは，インターネットの普及に伴う情報の量と質に十分対応可能な言語処理手法の一つを提案することにも繋がる．開発した分野語義データベース，及び学習手法は,情報検索をはじめ，質問応答やフィルタリングなど文書分類以外の様々なタスクに適用可能である．併せて，分野語義データベースの利用過程で生じる問題点は，現在の学習理論，言語処理技術，及び言語資源にも還元することができることから，本研究の学術的意義は極めて大きい．

研究成果の概要（英文）：The objective of this research is to develop lexical-semantic processing and supervised deep learning technique for text classification that classify test documents which are different period of the training documents. We have published five papers as our research output this pear. More specifically, (1) identification of domain-specific senses, (2) two approaches for word sense disambiguation, and (3) semantic equivalence. Furthermore, we proposed a method to utilize the lexical semantics of a label assigned to the training data and semantics of words in a document to improve the overall performance of text classification task. In future, we are going to extend our current domain-specific senses to fewl labeled training instances.

研究分野：自然言語処理

キーワード：分野依存語義　文書分類

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

文書分類に関する研究は，機械学習が提案された 90 年代初頭から教師付き学習による分類が主流となっている． 教師付き学習は高精度な分類が期待できる反面， 正解ラベル付き訓練文書の作成には多大な労力 とコストを要する． この問題に対処するため， 半教師付き学習 ［Dagan’95， Beygelzimer’09,Settles’10］ や 能動学習 ［Yarowsky’95, Blum’98， Jpachims’99， Blum’01， Kingma’14］， あるいは転移学習 ［Caruana’96, Daume’06， Dai’07］ などの学習手法が提案されている． しかし， 訓練文書と異なる時期に作成されたテスト文書を分類する場合のモデルの頑強性については， なお検討の余地がある．

教師付き機械学習を用いた分類手法の多くが語の表層情報を利用しているのに対し， 語義を考慮した初期の研究として，Yang や Allan らの研究がある［Yang’94，Allan’98]．彼らは， 辞書情報を用いた単語の類推や構文解析などの言語処理を利用し， 分類を試みた．しかし， いずれの場合にも大幅な精度の向上がみられなかったことから， 言語処理の利用は必ずしも有効ではないと結論づけている． 一方で， GPU をはじめとする計算機処理能力の飛躍的な進展， 及びビッグデータが利用できるようになったことを背景に， 近年， ニューラルネットワークが再び脚光を浴びており， 文書分類タスクにおいてもこれまでにない高い精度が得られることが次々に報告されている． さらに言語処理においては， 語に関する意味の分散表現 ［Mikolov’13］ をモデル学習時に取り入れることにより， 高い精度が得られることが報告されている［Zhang’15， Wang’15， Joulin’16]．しかし，本研究で対象としている訓練文書と作成時期が異なるテスト文書の分類精度は， ニューラルネットワークなどを用いても高精度な分類までは至っておらず， 依然として模索が続いている[Salles’13，Fukumoto’15]．

本研究は， 訓練データと作成時期が異なるテスト文書を高精度で分類するためには， 意味を中心に据えた自然言語処理技術が必要不可欠であるという主張のもとに， 長期間に及ぶ時系列テキストデータを対象とした分類に有効な語彙的意味処理技術と機械学習法を開発することを目的とする．

２．研究の目的

本研究の大目標は，高精度な文書分類を実現するために必要な語彙的意味処理と学習法を明らかにすることである．具体的には，(1) 各分野に対し，大域的な特徴量を持つ語義と分野に依存する語義を抽出する．さらに，語彙に関する意味処理において分類に貢献する適切な意味の粒度，及び異なる言語における分野語義の類似 ・相違点についても明らかにし，それらの知見を言語処理へ還元する．学習では (2) 文書に付与されている分野語義と文書との意味的な関係を利用する深層学習を提案することにより，高精度な分類を目指す．

３．研究の方法

本研究は 3 つの課題から成る．第 1 の課題は，語義の分野付与である．訓練文書を利用し，単語の各語義がどの分野で頻出するかを同定することにより分野語義データベース (DDB) を構築する．第 2 の課題は，大域的特徴量を持つ語義の抽出である．年度にかかわ

らず一貫して各分野を特徴づける語義を抽出する．第 3 の課題は，文書に付与されている分野語義と文書との意味的な関係を利用する深層学習を提案することにより，高精度な分類を目指す．

４．研究成果

本研究の成果を以下に示す．

(1) 分野依存語義の同定

分野依存語義に関して， 2 つの手法を提案した．一つ目は，単語埋め込み表現を利用し，PageRank を適用することにより，分野ごとに target word の語義を特定する手法を提案した．

PageRank で用いる各ノード(単語の語義)は単語の埋め込み表現により，ノード間の関係を Word Mover Distance と呼ばれる手法によりノード間の意味的な距離を求め，PageRank 手法を適用した[論文 1]．

2 つ目は，graph ベースの深層学習を用いることにより，分野に依存する語義を高精度で抽出する手法である．具体的には，Neural Random Walk Model を用い，各分野に依存する語義を効率良く学習する手法を提案し，分野依存語義が文書分類の精度に貢献することを示した[論文 2]．

(2) GCNs に基づく語義の曖昧性解消

局所的特徴量と大域的特徴量に注目する手法として， それぞれ深層学習 GCNs， 及び Transformer XL を用い学習することにより，高精度な語義の曖昧性解消が行えることを明らかにした．具体的には， 局所的な特徴である Local features を GCNs により学習し，局所 (Local features)、及び大域的特徴量である Global features を Transformer-XL を用い同時に学習する手法を提案した[論文 3]．

(3) 語義の等価性に関する同定

語義の等価性に関する同定として本手法ではパラフレーズを設定した．パラフレーズとは，意味が同値である2文のことであり，パラフレーズ判定は，入力文対の意味が同じか否かを判断するタスクである．意味の同値性を高精度で判定するためには，一文の意味はもちろんのこと，文外の情報，すなわち文間関係を捉える必要がある．本研究ではこの問題に対し，談話構造における文間関係の一つであるelaboration (精緻化)に注目した．具体的には、2文がelaboration の関係にある場合には，一方の文が他方の文の内容をさらに詳細に述べているため，同値性が存在しないと仮定し，パラフレーズを高精度で判定する手法を提案した[論文4]．

(4) 文書分類への適用

本研究は，Amazonなどの商品名や新聞記事のタイトルなど，短い単語列，かつマルチラベルを対象とし，少数から成るこれらラベル付き事例を予め設定された階層構造のディレクトリへ高精度で分類する手法を開発するタスクを設定した．これまで従来手法の多くがテキストから得られる情報のみを利用していたのに対し，本研究ではショートテキストとそれに付与された分野との間には包含的な意味関係が存在することに注目し，大量の教師なしデータから学習した分野の意味表現とテキストの意味表現との関係性を利用することにより高精度な分類を目指した．具体的には，Graph Attention Network に疎なデータで

も対応できるようAdditive Attention をDot-Product-Attentionに変更することにより，カテゴリ間の共起関係を高精度で抽出し，文書表現と合成を施すことにより高精度で分類できる手法を提案した.

(5) 局所・大域的特徴量に基づく語義の曖昧性解消

各分野に対し，大域的な特徴量，すなわち年度にかかわらず同一の語義を判定する手法を提案した．具体的には，前文や後文に含まれる重要なキーワードに着目，文書単位でキーワードを捉え語義を解消する手法を提案した．対象単語、対象単語が属する文と文書，及び対象単語の候補語義文を Transformer と呼ばれるニューラルネットワークを用いて埋め込む．またモデル訓練のために Cross-Entropy Loss を用いた．ベンチマークデータセットを用いた実験の結果，ALL のスコアが 1.5%向上していること，先に本研究で提案した GCNs に基づく手法[論文 3]と比較しても ALL で 9.5%と大幅に精度が向上していることから，手法の有効性を示すことができた[論文 5]

発表論文

[論文1] Attaporn Wangpoonsarp and Kazuya Shimura and Fumiyo Fukumoto, Unsupervised Predominant Sense Detection and its Application to Text Classification, Applied Sciences, Vol. 10. No. 4, 2020.

[論文2] Attaporn Wangpoonsarp and Fumiyo Fukumoto, Predominant Sense Acquisition with a Neural Random Walk Model, International Conference on Neural Information Processing, pp.51-59, 2021.

[論文3] Fumiyo Fukumoto, Taishin Mishima, Jiyi Li, and Yoshimi Suzuki, Neural Local and Global Contexts Learning for Word Sense Disambiguation, International Conference on Neural Information Processing, pp. 531-539, 2021.

[論文4] Sheng Xu, Fumiyo Fukumoto, Jiyi Li, Yoshimi Suzuki, Paraphrase Identification with Neural Elaboration Relation Learning, International Conference on Neural Information Processing, pp. 562-573, 2021.

[論文5] 浅川翔，鈴木良弥，李吉屹，福本文代，局所および大域的特徴量に基づく語義の曖昧性解消，第28回言語処理学会年次大会，pp. 1797-1801，2022.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　4件）

１．著者名 Sheng Xu, Xingfa Shen, Fumiyo Fukumoto and Jiyi Li	４．巻 10(12)
２．論文標題 Paraphrase Identification with Lexical, Syntactic and Sentential Encodings	５．発行年 2020年
３．雑誌名 Applied Sciences	６．最初と最後の頁 4144
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/app10124144	査読の有無 有
オープンアクセス 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 該当する

１．著者名 Fumiyo Fukumoto, Yoshimi Suzuki, Attaporn Wangpoonsarp, and Meng Ji	４．巻 1
２．論文標題 Integrating Local and Global Data View for Bilingual Sense Correspondences	５．発行年 2019年
３．雑誌名 Knowledge Discovery, Knowledge Engineering and Knowledge Management	６．最初と最後の頁 224-241
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 オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kazuya Shimura, Joyo Li, and Fumiyo Fukumoto	４．巻 1
２．論文標題 HFT-CNN Learning Hierarchical Category Structure for Multi-label Short Text Categorization	５．発行年 2018年
３．雑誌名 2018 Conference on Empirical Methods in Natural Language Processing	６．最初と最後の頁 811-816
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 F. Fukumoto and Y. Suzuki and A. Wangpoonsarp	４．巻 2
２．論文標題 Is (President, 大統領) a Correct Sense Pair? Linking and Creating Bilingual Sense Correspondences	５．発行年 2017年
３．雑誌名 9th International Conference on Knowledge Engineering and Ontology Development	６．最初と最後の頁 39-48
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0006489900390048	査読の有無 有
オープンアクセス 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 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　4件）

1．発表者名
Yoshimi Suzuki and Fumiyo Fukumoto

2．発表標題
Integrating Internet Directories by Estimating Category Correspondences

3．学会等名
KEOD（国際学会）

4．発表年
2019年

1．発表者名
Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto

2．発表標題
Text Categorization by Learning Predominant Sense of Words as Auxiliary Task

3．学会等名
Proc. of the 57th Annual Meeting of the Association for Computational Linguistics（国際学会）

4．発表年
2019年

1．発表者名
志村和也，李吉屹，福本文代

2．発表標題
Fine-tuningに基づくショートテキストの自動分類

3．学会等名
情報処理学会，第13回テキストアナリティクス・シンポジウム

4．発表年
2018年

1．発表者名
K. Shimura and F. Fukumoto

2．発表標題
Title Categorization based on Category Granularity

3．学会等名
Proc. of the 8th Language and Technology Conference（国際学会）

4．発表年
2017年

１．発表者名
A. Wangpoonsarp and F. Fukumoto

２．発表標題
Identification of Domain-Specific Senses based on Word Embedding Learning

３．学会等名
Proc. of the 9th International Conference on Knowledge Engineering and Ontology Development（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
福本・李研究室
cl.cs.yamanashi.ac.jp/index.html

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