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　６ 年　 ６ 月 １１ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２０３８

研究課題名（和文）Theoretically founded algorithms for the automatic production of analogy tests in NLP

研究課題名（英文）Theoretically founded algorithms for the automatic production of analogy tests in NLP

研究代表者

ＬＥＰＡＧＥ　ＹＶＥＳ（LEPAGE, YVES）

早稲田大学・理工学術院（情報生産システム研究科・センター）・教授

研究者番号：７０５７３６０８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：近年の人工知能で、単語や文の意味を数字で表現する。意味が正しく表現されるかを評価するため、類推データセットを用いる。しかし、類推データセットの構築は、今まで自動化されず、人手で英語で構築されたものは日本語に翻訳されても、英語へ偏り、さらに主に特別な種類の類推関係に偏っている。多言語の類推データセットを自動的に構築することで、規則・不規則の単語分解や生成に役に立つを示し、単語間の意味的な新しい類推関係の発見できた。文間類推データセットの構築より、どの文のパターンが類推関係をより多く含まれるかと理解した。言い換えに基づく文間類推データセット構築を提案し、類推関係を理解する神経回路モデルも提案した。

研究成果の学術的意義や社会的意義
人間の性質な認知行動の一つは、類推関係を認識することである。例えば、「男」:「女」::「王」:何？との質問には「妃」の答えは可能だ。また、「この曲は好き。」:「歌ういたい気分だ。」::「このゲームは好き。」:「プレーする気がする。」は文間の例になる。
最先端人工知能の単語や文の表現では、どの程度その認知能力を持っているか、それを測るために、類推関係データセットが必要とのなる。本研究では単語間と文間類推データセットの構築を検討した。英語だけでなく、多言語可能な手法、さらにある古典的な類推関係だけでなく（性別、国・首都）、より幅広い手法を提案と検討した。

研究成果の概要（英文）：　Recent artificial intelligence uses numbers to represent the meaning of words or sentences. In order to evaluate whether the meaning is correctly represented, analogy datasets are used. However, the construction of analogy datasets has not been automated until now, and those constructed manually in English are biased toward English, even when translated into Japanese, and biaised toward special types of analogical relations.
By automatically constructing multilingual analogical datasets, we were able to show that it is useful for regular and irregular word analysis and generation, and to discover new semantic analogical relations between words. From the construction of sentence analogy datasets, we understood which sentence patterns contain more analogical relations. We proposed a paraphrase-based sentence analogy dataset construction method, and also proposed neural circuit models for understanding/solving analogical relations.

研究分野：自然言語処理・人工知能

キーワード：認知能力　類推関係　類推関係の徹底的抽出　単語埋め込み空間　文間類推関係のための神経回路モデル　実数値間類推関係　ブール値間類推関係　整数値間類推関係

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景：

（１）近年の自然言語処理の分野では、単語や文のベクトル表現とその表現は最も重要な進捗であった。複数のタスクにおいて評価するのは外部的評価となる。主な内部的評価手法として、単語ベクトル表現の品質を測定することは、類推テストセットを用いて行う。

（２）しかし、類推データセット（アナロジーデータセット、４個類推データセット）の構築は、今まで自動化されていない。また、人手で英語で構築されたテストセットを日本語に翻訳しても、英語への偏りがあり、さらに特別（百科事典的な知識）な種類の類推関係にも偏っている。

２．研究の目的：

（１）ベクトル表現の本質的な評価を行うためのツールを装備する背景の下で、単語や文のベクトル表現空間から、類推データセットを自動的で徹底的に抽出するため、ツールの設計を検討し、その完全性や効率性を測る。

（２）同時に上記の2つの偏りを取り払うことを目標する。様々な言語での類推データセットを公開する目標もある。既存の自動的類推関係抽出手法が整数値ベクトル表現用の手法が存在するため、単語埋め込み空間や文表現の実数値に適応する可能性を検討した。

３．研究の方法：

実数値ベクトル空間内の類推関係を徹底的に抽出する課題に関して、以前開発した整数値ベクトル空間用のプログラムでは、編集距離を利用したため、編集距離を実数値ベクトルで表現できるかとの課題（研究課題(a)）を扱った上に、実数値ベクトル空間に拡張する可能性を検討した（研究課題(b)）。加速化し、整数値から実数値へ変換するには、実数値の近似あるいは文間の緩和などを研究した（研究課題(c)）。単語間も文間類推データセットも構築し、以前のプログラムを使用し、様々な神経回路手法を使用しても研究した。神経回路の実験を行うため、Deep LearningBoxとGPUカードを購入した。（研究課題(a)と(b)）のために博士課程学生を、プログラムの加速化のために修士学生を雇用した。

４．研究成果：

（１）値の種類の検討（研究課題(b)）を行う必要性は以下にある。研究代表者が、文字列間類推関係の研究で、整数値ベクトル表現に長らく取り組んできた。同時に、特に人工知能の分類機の研究する多くの研究者が、ブール値の特徴ベクトルが重要なものであると強調している。それに対して、単語埋め込み空間ベクトルや文ベクトル表現は実数値でのベクトルや行列であるため、実数値を扱えることはこの研究にとって重要である。

（1a）ベクトル並列計算ライブラリの導入により、以前に開発したプログラムを加速化がした（研究課題(c)）。また、非決定的アルゴリズムを実装し、10言語の大規模な形態素類推関係データテストにおいて、他の既存手法と比較し、最高の妥当性を示した[Deng and Lepage, ATA@ICCBR, 2023]。さらに、理論的な成果として、ブール値間類推関係方程式の解の新しい公式を公開した [Lepage, IARML@IJCAI, 2023]。

（1b）整数値から実数値表現への移動（研究課題(b)）の際、否定的結果をえられた：編集距離をベクトル空間に埋め込む手法は、理論的に不可能だ。いくつかの類推関係データセットで、いくつかの言語で、いくつかの単語埋め込み空間を使用し、様々な近似手法の検討の結果、類推関係が見出せられないという事実は実験で明らかにした[Fam and Lepage, LTC, 2023]。

（1c）上記の(1a)と(1b)の困難性をめぐった結果、期待できる新研究問題を提起した。実数間類推関係の性質の定義を一般化平均（別称ヘルダー平均）に基づいた正実数間の類推関係の定義を提案した[Lepage and Couceiro, JIAF, 2024]。

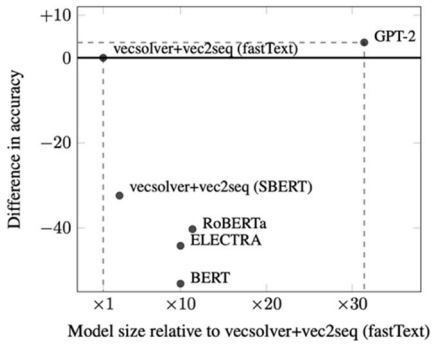
（２）単語間類推データセットの構築に当たって、形式と意味的のレベルを扱った。

（2a）百以上の言語での形態統語論素性付き語形のデータセットを利用し、高速化したプログラム（(1a) を参照）を活用し、自動的に類推関係グリッドと呼ぶデータ構造を全て抽出した。このようなデータ構造が類推データセットを要約することとみなせる。資源の少ない言語では、規則・不規則形態素解析と生成のタスクで効果的であると証明した[Fam and Lepage, JETAI, 2022]、[Fam and Lepage, AMAI, 2024]。

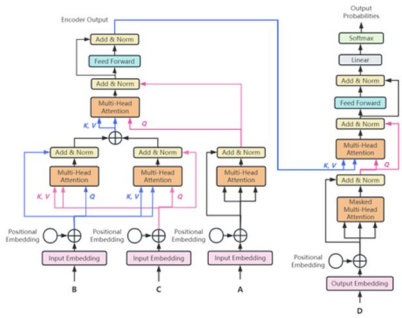
（2b）また、単語埋め込み空間から類推関係の徹底的抽出のための手法を提案した。二等分線の計算に基づき、類推関係クラスターを段階的に抽出し、交わりをとる手法である。すでに公開された類推データセットと評価した[執筆中・未発表]。

（３）文間類推データセットの構築に当たって、実数値ベクトル空間の探索の複雑さを控えるため（研究課題(b)、成果（1b））、異なる方面を探ることにした（研究課題(a)）。

（3a）まず、単語間類推関係に含まれている単語の定義文章で言語モデル学習を続けることで、ある意味タスクにおける言語モデルの質を向上させることを示した [Zhang et Lepage, ATA@ICCBR, 2023]。また、類推関係を近似することでテキスト含意・要因文対データセットを元にし、文間類推データセットを構築した。実験では、微調整より、以上の文間類推データセットの使用の方で、学習済み言語モデルの改良に至ると明らかにし、緩い類推関係が暗黙的だが一貫した規則性を効果的に捉えることができると示した[Wang et al., LTC,2023][Wang and Lepage, AMAI, 2023]。最後に、新しい意味形式的文間類推データセットの構築手法を提案した。既存単語間類推関係に基づき、逆翻訳により言い換えで得られた文に単語を入れ替え、同様の比例を持つ４つの文、すなわち類推関係、の生成する手法である。英語のみデータセットを構築したが、複数の言語に応用できる手法である。以前の手法に比べ、データセットに含まれる文がより長く、語彙がより豊かで、意味的に幅がより広くとの立点がある[Yan et al., IARML@IJCAI, 2024]。



（3b）データセットの構築のため、文間類推の解を求めるために類推関係の性質を元にした最先端神経回路モデルのトランスフォーマーモデルを提案した。以前の手法に比べ、より複雑セットで優れていると明らかにした[Yan et al., IARML@IJCAI, 2024]。



（3c）与えられたデータセットに含まれる類推関係の量を測るため、類推関係密度という尺度を提案した。様々な言語において、間類推関係の量を測り、より高い密度の文パターン粒度を特定した[Fam and Lepage, Information, 2023]。

＜引用文献＞

以上の研究成果で挙げた引用文献は全て査読付きワークショップや国際会議のものである。以下の「主な発表論文」をご参照。

５．主な発表論文等

〔雑誌論文〕　計15件（うち査読付論文　15件／うち国際共著　2件／うちオープンアクセス　11件）

１．著者名 M. Eget, X. Yang, and Y. Lepage	４．巻 －
２．論文標題 A study in the generation of multilingually aligned middle sentences	５．発行年 2023年
３．雑誌名 Proceedings of the 10th Language & Technology Conference (LTC 2023) & Human Language Technologies as a Challenge for Computer Science and Linguistics	６．最初と最後の頁 45-49
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 R. Fam and Y. Lepage	４．巻 －
２．論文標題 Investigating parallelograms: Assessing several word embedding spaces against various analogy test sets in several languages using approximation	５．発行年 2023年
３．雑誌名 Proceedings of the 10th Language & Technology Conference (LTC 2023) & Human Language Technologies as a Challenge for Computer Science and Linguistics	６．最初と最後の頁 68-72
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 L. Wang, Z. Pang, H. Wang, X. Zhao, and Y. Lepage	４．巻 －
２．論文標題 Solving sentence analogies by using embedding spaces combined with a vector-to-sequence decoder or by fine-tuning pre-trained language models	５．発行年 2023年
３．雑誌名 Proceedings of the 10th Language & Technology Conference (LTC 2023) & Human Language Technologies as a Challenge for Computer Science and Linguistics	６．最初と最後の頁 325-330
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 Fam Rashel、Lepage Yves	４．巻 －
２．論文標題 Organising lexica into analogical grids: a study of a holistic approach for morphological generation under various sizes of data in various languages	５．発行年 2022年
３．雑誌名 Journal of Experimental & Theoretical Artificial Intelligence	６．最初と最後の頁 1-26
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/0952813X.2022.2078890	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 Fam Rashel、Lepage Yves	4．巻 12
2．論文標題 A Study of Analogical Density in Various Corpora at Various Granularity	5．発行年 2021年
3．雑誌名 Information	6．最初と最後の頁 314～314
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/info12080314	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 Wang Liyan、Lepage Yves	4．巻 －
2．論文標題 Learning from masked analogies between sentences at multiple levels of formality	5．発行年 2023年
3．雑誌名 Annals of Mathematics and Artificial Intelligence	6．最初と最後の頁 1--25
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10472-023-09918-2	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 Fam Rashel、Lepage Yves	4．巻 －
2．論文標題 A study of universal morphological analysis using morpheme-based, holistic, and neural approaches under various data size conditions	5．発行年 2024年
3．雑誌名 Annals of Mathematics and Artificial Intelligence	6．最初と最後の頁 1--25
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10472-024-09944-8	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 X. Deng and Y. Lepage	4．巻 3438
2．論文標題 Resolution of analogies between strings in the case of multiple solutions	5．発行年 2023年
3．雑誌名 In CEUR, editor, Proceedings of ICCBR: Workshop on Analogies: from Theory to Applications (ATA@ICCBR 2023), CEUR Workshop Proceedings	6．最初と最後の頁 3--14
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 Y. Lepage	4．巻 3492
2．論文標題 Formulae for the solution of an analogical equation between Booleans using the Sheffer stroke (NAND) or the Pierce arrow (NOR)	5．発行年 2023年
3．雑誌名 In M. Couceiro, P.-A. Murena, and S. Afantenos, editors, Proceedings of the Workshop Interactions between analogies and machine learning, co-located with IJCAI 2023 (IARML@IJCAI 2023)	6．最初と最後の頁 3--14
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 Y. Lepage and M. Couceiro	4．巻 －
2．論文標題 Analogie et moyenne generalisee	5．発行年 2024年
3．雑誌名 Actes de la conference Journees d ' intelligence artificielle francaises & Plateforme francaise d ' intelligence artificielle (PFIA-JIAF 2024)	6．最初と最後の頁 －
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 L. Wang, Z. Pang, H. Wang, X. Zhao, and Y. Lepage	4．巻 －
2．論文標題 Solving sentence analogies by using embedding spaces combined with a vector-to-sequence decoder or by fine-tuning pre-trained language models	5．発行年 2023年
3．雑誌名 In Z. Vetulani and P. Paroubek, editors, Proceedings of the 10th Language & Technology Conference (LTC 2023) -- Human Language Technologies as a Challenge for Computer Science and Linguistics	6．最初と最後の頁 325--330
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 L. Wang, H. Wang, and Y. Lepage	4．巻 －
2．論文標題 Continued pre-training on sentence analogies for translation with small data	5．発行年 2024年
3．雑誌名 Proceedings of the 14th International Conference on Language Resources and Evaluation (LREC 2024) and the 30th International Conference on Computational Linguistics (COLING ' 24)	6．最初と最後の頁 3890--3896
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 B. Yan, H. Wang, L. Wang, Y. Zhou, and Y. Lepage	4．巻 —
2．論文標題 Transformer-based hierarchical attention models for solving analogy puzzles between longer, lexically richer and semantically more diverse sentences	5．発行年 2024年
3．雑誌名 In M. Couceiro, P.-A. Murena, and S. Afantenos, editors, In Proceedings of the Workshop Interactions between analogies and machine learning, co-located with IJCAI 2024 (IARML@IJCAI 2024)	6．最初と最後の頁 —
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 Y. Lepage and M. Couceiro	4．巻 —
2．論文標題 Towards a unified framework of numerical analogies: Open questions and perspectives	5．発行年 2024年
3．雑誌名 In M. Couceiro, P.-A. Murena, and S. Afantenos, editors, In Proceedings of the Workshop Interactions between analogies and machine learning, co-located with IJCAI 2024 (IARML@IJCAI 2024)	6．最初と最後の頁 —
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 Q. Zhang and Y. Lepage	4．巻 3438
2．論文標題 Improving sentence embedding with sentence relationships from word analogies	5．発行年 2023年
3．雑誌名 Proceedings of ICCBR: Workshop on Analogies: from Theory to Applications (ATA@ICCBR 2023), CEUR Workshop Proceedings	6．最初と最後の頁 43--53
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 —

〔学会発表〕　計8件（うち招待講演　3件／うち国際学会　1件）

1．発表者名 Yves Lepage
2．発表標題 Giving a structure to language data: from analogies to analogical grids.
3．学会等名 Invited talk at the seminar of Dublin City University (DCU), 4th of July 2022. (招待講演)
4．発表年 2022年

１．発表者名
Yves Lepage

２．発表標題
Analogy on text data

３．学会等名
Invited talk at the workshop Interaction between Analogical Reasoning and Machine Learning (IARML 2022), 23rd of July 2022.（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
Y. Lepage

２．発表標題
Analogie et donnees de langue

３．学会等名
Colloquium LORIA, Nancy, 15 novembre 2023, LORIA, https://www.loria.fr/fr/2023/10/colloquium-loria-yves-lepage/.（招待講演）

４．発表年
2023年

１．発表者名
Y. Lepage

２．発表標題
Analogie, explication des donnees de langue et travaux recents sur representations vectorielles de phrases et analogie

３．学会等名
Workshop Analogies: From learning to explainability, Arras, 27--28 novembre 2023, https://www.loria.fr/event/workshop-analogies-from-learning-to-explainability-with-zied-bouraoui/

４．発表年
2023年

１．発表者名
Y. Lepage

２．発表標題
Analogie et moyenne : considerations generales et application aux chaines

３．学会等名
Forum sciences cognitives et traitement automatique des langues, Nancy, 29 nov. 2023, https://idmc.univ-lorraine.fr/evenements-idmc/fsc-2023/

４．発表年
2023年

１．発表者名
Y. Lepage

２．発表標題
Jeux d'analogies pour le TAL

３．学会等名
MALOTEC/LORIA seminar, Nancy, 13 Dec. 2023, https://malotec.loria.fr/

４．発表年
2023年

１．発表者名
Y. Lepage

２．発表標題
Analogie et moyenne generalisee (Analogy and generalized means)

３．学会等名
MALOTEC/LORIA seminar, Nancy, 14 Feb. 2024, https://malotec.loria.fr/

４．発表年
2024年

１．発表者名
Y. Lepage

２．発表標題
Jeux d'analogies pour le TAL

３．学会等名
RALI-OLST seminar, Montreal, 12 June 2024, http://rali.iro.umontreal.ca/rali/?q=fr/node/1222/list/2024

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
Kakenhi Kiban C 21K12038
http://lepage-lab.ips.waseda.ac.jp/projects/Kakenhi_Project_21K12038/
Tab: Experimental Results

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
研究協力者	ファム　ラシェル (FAM RASHEL)	早稲田大学・大学院情報生産システム研究科・博士後期課程	類推関係に基づく形態素解析・生成、文間パターン類推関係の研究
研究協力者	王 (WANG LIYAN)	早稲田大学・大学院情報生産システム研究科・博士後期課程	文間類推関係のための神経回路モデル
研究協力者	フイドロム　ルダリ (HUIDROM RUDALI)	早稲田大学・大学院情報生産システム研究科・修士学生	形式類推関係の徹底的抽出プログラムの高速化
研究協力者	コセイル　ミゲル (COUCEIRO MIGUEL)	ロレーヌ大学・LORIA・教授	実数間類推関係の数学的研究

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			
フランス	ロレーヌ大学	LORIA		