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　３ 年　 ６ 月　 ４ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１４４７

研究課題名（和文）Self-explainable and fast-to-train example-based machine translation using neural networks

研究課題名（英文）Self-explainable and fast-to-train example-based machine translation using neural networks

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）：類推関係に基づいた用例機械翻訳システムに、自己説明を導入した。従って、本研究は説明可能な人工知能（XAI）に位置づけられる。翻訳過程の自己説明機能としては、文を再帰的に翻訳する際、既にある翻訳メモリーから似た文を検索し、翻訳した際、どういうふうに使用したか、そのトレースを翻訳システムに導入した。与えられた文を意味的にも形式的にもカバーする文を検索する手法を開発した。文間類推関係のデータセットを公開した。コーパスの類推関係密度を調べた。

研究成果の学術的意義や社会的意義
統計的機械翻訳（SMT）の手法である文部分アライメントと、ニューラル自然言語処理（NMT）の手法である単語や文のベクトル表現を用いて、類推関係方程式の解を求める手法を改善した。類推関係に基づいた用例機械翻訳の直接的なアプローチと間接的なアプローチを融合し、独自のニューラルネットワークを用いたシステムを構築した。入力は、単語のベクトル表現に基づく、単言語または対言語のソフトアライメントです。

研究成果の概要（英文）：This research introduced self-explanation in example-based machine translation (EBMT) by analogy. It is thus positioned in explainable artificial intelligence (XAI). Self-explanation was implemented by tracing the analogies verified or solved during translation. The direct and indirect approaches to EBMT by analogy were merged in system that uses an original neural network. It was studied how to retrieve sentences that cover a given sentence semantically and formally was built. It was studied how dense corpora are relative to analogies. Datasets of analogies between sentences were released.

研究分野：自然言語処理

キーワード：自然言語処理　機械翻訳

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

（１）翻訳者は翻訳メモリ（TM）を使って作業を行うが、TM の主な機能としては、翻訳するべき文に似た文を検索する機能です。検索された文は、翻訳者が作成する翻訳結果の説明になる。本研究では、機械翻訳に自己説明機能を、いわば、文はどういうふうに翻訳されたかの説明機能を導入することについての検討となる。そのため、本研究は、説明可能な人工知能（XAI）に位置づけられる。

（２）用例機械翻訳（EBMT）、統計的機械翻訳（SMT）、ニューラル機械翻訳（NMT）は、データに基づいた機械翻訳の手法である。SMT 及び NMT は、対訳コーパスから学習（訓練）を行うため、長時間が必要となる。代わりに翻訳の際早い。しかし、特にニューラル機械翻訳の場合、翻訳の説明はできない。不透明であるため、ブラックボックスだという。EBMT は遅延学習の手法であり、基本的に事前に知識抽出過程（訓練）がなく、翻訳する際データを活躍させ、翻訳の際時間がかかる。EBMT は TM のように類似文を検索し、翻訳過程の自己説明を提供する可能性がある。

２．研究の目的

（１）本研究の目的は、自己説明機能を持ち、訓練時間の短い EBMT システムを設計することであった。翻訳過程の自己説明機能としては、文を再帰的に翻訳する際、そのトレースを表示する。訓練時間については、設計したシステムの訓練時間を測定または推定することにした。

（２）本研究の重要な課題として、SMT と NMT の効率的な技術を、設計する EBMT システムに貢献できるかとのことであった。特に、類推関係に基づいた用例機械翻訳システムには、SMT や NMT やニューラル技術を用いた自然言語処理（Neural NLP）の効率的な技術を使用することによって、類推関係の計算手法に貢献できるかとの検討した。

３．研究の方法

（１）本研究では、主に三つの次の技術を組み合わせることによって事項説明機能を持った類推関係に基づいた用例機械翻訳システムを設計した：（a）Neural NLP からの重要な学習技術の成果であり単語分散表現（分布意味論に基づき単語ベクトル表現又は単語埋め込みと呼ぶ）と、（b）SMT からの技術であり文部分アライメントと、（c）EBMT からの技術であり翻訳説明となる類推関係の再帰的翻訳のトレース。

（２）重要な研究課題としては、どういうふうに単言語的にも、二か国語的にも、連続的単語表現のソフトなアラインメントを導入するかと、類推関係に基づいた機械翻訳の枠にその応用できるかとのことであった。最先端技術の文ベクトル表現の導入についての研究も行なった。新しい手法を提案し、それに基づき、文間類推関係データを構築し、公開した。また、提案した手法がどの程度に適用できるかを検討し、コーパスの類推関係密度に関する研究を行った。

（３）物品費：提案した神経回路モデルの学習のため、機械学習専用の計算機（DeepLearning Box）を購入した。価格の高騰のため、予告通り二台目を購入することは断念し、その代わり、購入した一台の計算機に複数の GPU プロセッサを搭載した。資源に関して、当初検討していた BTEC コーパスの購入は、高価のため、断念した。代わりに、無料でダウンロードできる Tatoeba コーパスという多言語コーパスを使用した。また、文間類推関係のデータセットを構築した（下記研究成果（１）参照）。

（４）人件費：二人の研究補助者を採用した。一人は、文間類推関係方程式の解を求めるための神経回路モデルの設計と実装に取り組んだ（下記研究成果（1.b）及び（2.b）参照）。もう一人は、類推関係密度により、コーパス特徴付けの研究に取り組んだ（下記研究成果（4.c）参照）。

４．研究成果

（１）**データの使用とデータセットの構築。**
（a） 数ヶ国語の大規模単語埋め込みをダウンロードし、クリーンアップした。二ヶ国語の単語

埋め込みマッピングの新しい手法を提案し、実験を行った（国際会議 LTC 2019 で発表）。
（b） 英文の間の 5,000 以上意味と形式的な類推関係のデータセットを公開した。意味と形式的な類推関係方程式の解を求めるアルゴリズムを用いて構築したものである（以下（2.b）参照）。同様、英文の間の 5,000 以上の意味的類推関係のもう一つのデータセットを構築した。前者と同様なものであるが、文ベクトル表現を使用する提案した新しい神経回路モデルで構築されたものである（下記（2.b）参照）。両者のデータセットは本研究室のサイトで公開した。
（c） Tatoeba コーパスから対訳文間形式類推関係のデータセットを抽出した。データセットとして公開する予定。

（2）基礎アルゴリズムと基礎手法の開発。
（a）類推関係に基づいた用例機械翻訳システムに必要な構成要素であるため、与えられた文を形式的にも意味的にもカバーする文の集合を検索する構成要素を実装した。（i）単語ベクトル表現（Neural NLP の成果）と（ii）アライメント手法（SMT の成果）の技術を使用している。ベクトル空間におけるベクトル検索と、文字列のパターンマッチングを組み合わせたものである（国際会議で論文を提出予定）。
（b）文間類推関係方程式の解を求めるため、形式的アプローチと意味的アプローチの融合を提案した。形式的な類推関係を保持する文字列形式変換の理論的研究を行った（国際会議 ICACSIS 2018 で発表、学会最優秀論文賞受賞）。単語ベクトル表現（Neural NLP の成果）を用い、文間意味的・形式的類推関係方程式の解を求めるアルゴリズムを設計し、それを用いてデータを構築した（国際会議 LTC 2019 で発表）。 また、文ベクトル表現（NMT や Neural NLP の成果）を用いて文間類推関係方程式の解を求める新しい神経回路モデルを設計し、それを用いてデータを構築した（国際会議 ICACSIS 2019 で発表）。両者のデータセットを公開した。
（c）単言語又は二ヶ国語及びソフト又はクリスピー（前者は連続値、後者は０か１かの値のみ）な文間アライメントの効率性を検討した。新しい独創性のある類推関係に基づいた用例機械翻訳モデルを提案した。単言語で単語類似度測定には単語埋め込みを利用し（Neural NLP の成果）、対訳単語類似度の測定には、文部分アライメントを利用している（SMT の成果）。SMT で使用される対訳文アライメントを、（i）単語埋め込みを用いたクリスピーなアライメントからソフトなアライメントへ拡張し、（ii）単語翻訳確率を用いた対訳アライメントから単言語アライメントへ拡張した（下記（4.c）参照）。
（d）類推関係に基づいた用例機械翻訳システムの翻訳過程自己説明は、システムが選択した類推関係を表示することで説明する。インタフェースでのトレースの表示を実装した（ICCBR 2019 の論文で図を参照）。

（3）コーパス内の類推関係数の測定。類推関係密度に関して実験を行なった。類推関係数の多いコーパスと少ないコーパスを特徴つけるため検討した。実験の結果、出現頻度の低いサブワードをマスクすることにより、類推関係密度を高めることができると明らかにした。また、文の長さによって、翻訳がしやすくなる可能性があるため、文が短いものと長いものの複数のコーパスを使用し、類推関係密度を測定した（雑誌論文を投稿する予定）。

（4）類推関係に基づいた用例機械翻訳（EBMT）システムの実装。直接法と間接法を用いた二つの機械翻訳システムを実装した（以下（a）と（b））。また、その二つの手法の統合について検討した（以下（c））。
（a）第一目の機械翻訳システムは、直接法を採用している。そこで修正と検索には数値的な手法を導入した（国際会議 ICACSIS 2018 で発表）。
（b）第二目の機械翻訳システムでは、間接的なアプローチを採用している（国際会議 ICCBR 2018 で発表）。翻訳の過程を事例ベース推論（検索、再利用、修正、記憶の四段階のプロセスに基づく推論手法）の枠で形式化した（国際会議 ICCBR 2019 で発表）。その第二目の機械翻訳システムでは、実行トレースの表示により、すなわち原言語と目的言語での類推関係を表示することにより、翻訳過程の自己説明機能を実現した。類推関係に基づいた用例機械翻訳システムでは、類推関係が基本的な演算であるため、単語間の比較結果の種類（ソフト又はクリスピー）によってどの程度類推関係が立てられるかとその関係の解を求められるかを検証した。その第二目の機械翻訳システムでは、訓練時間は短くて（事前に学習したモデルを使用する場合は不要）、翻訳にかかる時間は長くなる。
（c）翻訳を行うため、ソフトな単言語文アライメントと対訳文アライメントを利用した新しい神経回路モデルを提案と実装したことによって、直接法（上記（a）参照）と間接法（上記（b）参照）を融合した。二ヶ国語（現言語と目的言語）と単言語で（目的言語）同時に成立した複数の類推関係方程式の解を求める手法である。事前に訓練された単語埋め込みを利用し、文間の対応関係（単言語の場合）または対訳関係（二か国語の場合）を表現するには、類似度行列を用いる。文部分アライメントと対訳単語埋め込みマッピングの使用と比較した（フランス自然言語処理学会大会で発表、会議最優秀論文賞受賞）。

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　9件／うち国際共著　1件／うちオープンアクセス　7件）

１．著者名 Lepage Yves、Lieber Jean	４．巻 11680
２．論文標題 An Approach to Case-Based Reasoning Based on Local Enrichment of the Case Base	５．発行年 2019年
３．雑誌名 Lecture notes in computer science (LNCS)	６．最初と最後の頁 235～250
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-29249-2_16	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 Lepage Yves、Lieber Jean	４．巻 11156
２．論文標題 Case-Based Translation: First Steps from a Knowledge-Light Approach Based on Analogy to a Knowledge-Intensive One	５．発行年 2018年
３．雑誌名 Lecture notes in computer science (LNCS)	６．最初と最後の頁 563～579
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-01081-2_37	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 He Kun、Zhao Tianjing、Lepage Yves	４．巻 -
２．論文標題 Numerical Methods for Retrieval and Adaptation in Nagao’s EBMT model	５．発行年 2018年
３．雑誌名 In Proceedings of the 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2018)	６．最初と最後の頁 195～200
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICACSIS.2018.8618226	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 Lepage Yves	４．巻 -
２．論文標題 String Transformations Preserving Analogies	５．発行年 2018年
３．雑誌名 In Proceedings of the 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2018)	６．最初と最後の頁 189～194
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICACSIS.2018.8618162	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 Vector-to-Sequence Models for Sentence Analogies	5．発行年 2020年
3．雑誌名 In IEEE, editor, Proceedings of the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2020)	6．最初と最後の頁 441～446
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICACSIS51025.2020.9263191	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 S. Yu and Y. Lepage	4．巻 -
2．論文標題 Iterative training for unsupervised word embedding mapping	5．発行年 2019年
3．雑誌名 Proceedings of the 9th Language & Technology Conference (LTC 2019). ISBN: 978-83-65988-31-7	6．最初と最後の頁 62～66
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 Lepage Yves	4．巻 -
2．論文標題 Semantico-formal resolution of analogies between sentences	5．発行年 2019年
3．雑誌名 Proceedings of the 9th Language & Technology Conference (LTC 2019). ISBN: 978-83-65988-31-7	6．最初と最後の頁 57～61
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 T. Zhao and Y. Lepage	4．巻 -
2．論文標題 Context encoder for analogies on strings	5．発行年 2018年
3．雑誌名 32th Pacific Asia Conference on Language, Information and Computation (PACLIC 32) URL: https://www.aclweb.org/anthology/Y18-1096/	6．最初と最後の頁 832～840
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 V. Taillandier, L. Wang et Y. Lepage	4．巻 2
2．論文標題 Reseaux de neurones pour la resolution d'analogies entre phrases en traduction automatique par l'exemple	5．発行年 2020年
3．雑誌名 In Actes de la 6e conference conjointe JEP-TALN-RECITAL. AFCP et ATALA. HAL-ID: hal-02784759	6．最初と最後の頁 108～121
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

〔学会発表〕　計1件（うち招待講演　1件／うち国際学会　1件）

1．発表者名 Y. Lepage
2．発表標題 An old but lively notion in AI and NLP: Analogy
3．学会等名 Keynote speech at ICACSIS 2020. URL: https://icacsis.cs.ui.ac.id/front/?page_id=1935（招待講演）（国際学会）
4．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Kakenhi Kiban C 18K11447 -- Results of experiments http://lepage-lab.ips.waseda.ac.jp/en/projects/kakenhi-kiban-c-18k11447/

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