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　 ５ 月 ２３ 日現在

機関番号：５７１０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ００８０６

研究課題名（和文）意味構造と統語構造の統合学習を用いた文書作成支援アプリケーションの開発

研究課題名（英文）Development of Writing Support Software Using Integrated Learning of Semantic and Syntactic Structures

研究代表者

小田　幹雄（Oda, Mikio）

久留米工業高等専門学校・制御情報工学科・教授

研究者番号：８０３００６４８

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：統語構造学習モデルと意味構造学習モデルを統合した言語モデルを構築し，第二言語習得者による文章作成の支援を目的として，文法と密接に関係する統語構造と意味構造学習モデルによる文法誤り訂正能力を明らかにした．また，言語コーパスにより訓練される文法誤り訂正モデルについて，人工訓練データを再構成する学習法を提案し，文法誤り訂正能力を向上させることに成功した．
さらに，英文を入力すると，提案した文法誤り訂正モデルにより文法誤りを指摘し，同時に，Wordnetデータベースから単語の概念と同義語を表示するアプリケーションソフトウエアを開発した．この成果は，第二言語学習者の学習支援に寄与するものである．

研究成果の学術的意義や社会的意義
これまで統語構造学習モデルと意味構造学習モデルを統合した言語モデルを検討した研究はなく，統合言語モデルを構築し，文法誤り訂正能力にどれくらい寄与するかを明らかにした．また，第二言語学習者を支援する文法誤り訂正モデルのための言語コーパスの改良法に関する本研究の手法は，文法誤り訂正の分野に限らず，教師あり学習モデルの構築に用いられる言語コーパスの改良に関する知見を与え，学習モデルの能力向上のために本手法が広く応用できると考えられる．さらに，統語構造と意味構造を表示するアプリケーションソフトウエアの試作は，今後，第二言語学習者の学習支援に寄与していくものと考えられる．

研究成果の概要（英文）：We examined a language model that integrates syntactic and semantic structure learning models, aiming to support sentence composition by second language learners. We elucidated the model's grammatical error correction capabilities, which are closely related to syntactic and semantic learning, which are closely related to grammar. Additionally, we proposed a learning method to reconstruct synthetic training data for grammatical error correction models trained on linguistic corpora, successfully improving their grammatical error correction capabilities.
Furthermore, we developed application software that, upon inputting an English sentence, identifies grammatical errors using the proposed grammatical error correction model and simultaneously displays the concepts and synonyms of words from the Wordnet database. These achievements contribute to supporting the learning of second language learners.

研究分野：自然言語処理

キーワード：言語モデル　文法誤り訂正　第二言語習得

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)言語処理には、おもに、データに基づく帰納的な方法論をとるコーパス言語学、計算機の膨大な記憶と処理能力による計算言語学の研究分野があり、計算言語学は、ニューラルネットワークの深層学習により、広範囲の応用分野において発展している過程にあった。言語における文の構成は、統語構造と意味構造からなるが、計算言語学による機械学習の応用分野では、おもに、統語構造に関係する言語コーパスを用いて学習する手法が一般的であり、統語構造と意味構造を学習する統合モデルの研究は十分ではなく、両者を統合する学習モデルを実現することにより、大きな技術的発展の可能性を秘めていた。とくに、意味構造の表現法の一つである抽象意味表現(Abstract Meaning Representation、 AMR）の適用可能性については、十分研究されていなかった。

(2)統語構造の一つである文法規則は、第二言語(L2)習得者が理解すべき規則の一つで、一方、言語処理分野においても、文法誤り訂正学習(GEC)モデルの研究は、CoNLL2014 Shared Task や BEA2019 Shared Task を経て発展してきた。しかしながら、これまでの学習モデルは、The Common Crawl corpus や Wikipedia 等の統語構造に基づく言語モデルと文法誤りを含む平文と文法誤りを訂正された平文の対を用いた教師あり学習モデルが主流であった。

(3)言語処理技術の教育への応用については、Workshop on Innovative Use of NLP for Building Educational Applications 等の国際会議でおもな研究成果が発表され、英文技術文書の作成支援が応用分野の一つである。目覚ましいグローバル化の発展に伴い、グローバルなコミュニケーションによる経済活動や学術活動の進展が望まれ、英文文書を作成する要請が高まっていた。英語の母語話者でない者が、第二言語である英語による文書を作成するとき、文書作成支援アプリケーションソフトウエアが望まれていた。

２．研究の目的

(1)確率的言語モデルに基づく言語処理の研究は、コーパスとそのコーパスに基づく単語の発生確率を利用した機構とその学習法を研究対象としており、 おもな研究課題は、学習機構の改良のほか、大規模な人工合成データを生成し学習に用いる、スペルチェッカーなどの単語情報を用いるなどがある。統語構造の一つである文法規則に関して、文法の誤りを訂正する GEC モデルがあり、このモデルは、言語習得者の第二言語文章作成を支援することができる。従来の文 GEC モデルを改良し、訂正能力を向上させ、母語話者と同程度の能力に近づけることを目的とする。

(2)言語習得者の L2 文章作成を支援する文法誤り訂正モデルについて、統語構造の学習モデルと意味構造の学習モデルを統合した GEC モデルを提案し、意味構造と統語構造を同時並列に学習することにより、文法誤り訂正能力の向上を図ることを目的とする。

(3)言語習得者の L2 文章作成を支援するために、作成した文章に対して、統語構造情報および意味構造情報を提供するアプリケーションソフトウエアを開発することを目的とする。

３．研究の方法

(1)統語構造の一つである文法構造を学習モデルで獲得するために、BEA2019 Shared Task の主要な研究成果で用いられている Transformers や Bidirectional Encoder Representations from Transformers の派生モデルに基づく教師あり学習モデルを使用し、Wikipedia 等による統語構造を学習した言語モデルを事前学習モデルとして用いる。この事前学習モデルを拡張して、文法誤りを訂正する学習モデルを構成する。学習には、文法誤りを含む文と正しい文との対からなる訓練データを用いて事後学習をする。ここで訓練データとして、BEA2019 Shared Task 等で公開されている L2 学習者の誤り文を母語話者が訂正したデータ(母語話者による注釈付きデータ)を用いる必要があるが、データの規模が小さいため、アルゴリズムにより誤り文を生成した人工合成訓練データも同時に用いられる。一般に、人工合成訓練データは、母語話者による注釈付きデータより品質が悪いため、人工合成訓練データによる訓練後のモデルの誤り訂正能力は低い。人工合成訓練データの品質を向上させる手法を検討し、L2 学習者にとって役立つ GEC モデルを実現する。なお、GEC モデルの性能は、m2 scorer 等を用い、$F_{0.5}$スコアで評価する。

(2)GEC モデルを発展させるために、(1)の統語構造を学習するモデルに意味構造を学習する Transformers または BERT の派生学習モデルをサブモジュールとして統合し、誤り訂正能力が向上するかを検討する。意味構造を学習するサブモジュールは、AMR による概念意味表現を入力とし、統語構造の文法的な誤りに対して、意味構造の情報を追加で付加することにより、文法誤り訂正能力が向上するかを検討する。なお、AMR を入力するサブモジュールは、1 次元トークン列の入力を前提とするため、グラフから 1 次元トークン列への変換法を検討する。

(3)L2 学習者の文書作成を支援するためのアプリケーションソフトウエアを開発する。L2 学習者が L2 の文(英文)を入力すると、研究方法(1)または(2)の文法誤り訂正モデルで統語構造で

ある文法誤りを指摘する機能を実装する。さらに、意味構造に関わる AMR や Wordnet データベースから文や単語の概念を表示する機能を実装する。

４．研究成果

(1)統語構造の一つである文法構造を学習した GEC モデルとして、Omelianchuk らの GECToR を用いた。従来の学習で用いられる母語話者の人手によりアノテーションされた L2 学習者コーパスは、データ量が少なく、十分な学習ができないため、GEC モデルの学習は、大規模な人工合成訓練データと組み合わせるパイプライン学習が行われてきた。これは、L2 学習者コーパスは高品質であるがデータ量が少ない、人工合成訓練データは低品質であるがデータ量が多いというトレードオフを解消するためである。訓練データは、文法誤りを含む文と文法的に正しい文の対から構成され、人工合成訓練データ生成アルゴリズムは、wikipedia 等からの文法的に正しい文にトークン的ノイズを挿入することにより、文法誤りを含む文を生成する。本研究は、生成された文法誤りを含む文に対して、元の文法的に正しい文が、必ずしも文法誤り文の正解文とはならないことを明らかにした。さらに、人工合成データを再構成することにより、L2 学習者コーパスと同程度にその品質を向上させる手法を提案した。再構成の手法は、つぎのとおりである。人工合成訓練データにおいて、既存の GEC モデルを用いて文法誤りを含む文から正しい文を推定し、この推定文を文法的に正しい文として置き換える。表 1 は、提案手法により再構成された人工合成訓練データを用いて学習した GEC モデルの誤り訂正能力を明らかにしたものである。なお、人工合成訓練データとして、しばしば利用される PIE(Parallel Iterative Edit Models for Local Sequence Transduction)および C4(Colossal Clean Crawled Corpus)を用いた。また、文法誤り訂正能力の評価は、CoNLL2014 および BEA2019 Shared Task の評価用データに対する$F_{0.5}$スコアとした。実験結果より、提案手法により再構成した人工合成訓練データのみによる学習法(Proposed)は、再構成前の人工合成訓練データのみによる従来の学習法より、かなり高い文法誤り訂正能力を実現できることがわかった。また、L2 学習者コーパスも用いて学習する従来のパイプライン学習(Original+BEA2019)と同程度の文法誤り訂正能力を獲得できることがわかった。

表１　人工合成訓練データとその再構成による誤り訂正能力の比較

Synthetic	Training Datasets	Stage			CoNLL2014 test			BEA2019 test		
		I	II	III	P	R	$F_{0.5}$	P	R	$F_{0.5}$
PIE-9M	Original	$\mathcal{S}$			60.4	31.8	51.2	54.3	41.5	51.1
	PartiallyRecovered	$\mathcal{S}$			58.7	32.3	50.4	51.2	42.9	49.3
	Proposed	$\mathcal{S}$			66.5	46.3	**61.2**	68.3	60.9	**66.7**
	Original+BEA2019	$\mathcal{S}$	$\mathcal{A}$		64.2	45.3	**59.2**	61.3	58.5	**60.7**
	PartiallyRecovered+BEA2019	$\mathcal{S}$	$\mathcal{A}$		63.7	45.5	59.0	60.0	59.3	59.8
	Proposed+BEA2019	$\mathcal{S}$	$\mathcal{A}$		64.0	45.1	59.1	60.0	58.8	59.8
	Original+BEA2019	$\mathcal{S}$		$\mathcal{A}$	72.4	40.2	62.4	76.4	53.4	70.3
	PartiallyRecovered+BEA2019	$\mathcal{S}$		$\mathcal{A}$	72.8	39.1	62.1	76.4	52.3	70.0
	Proposed+BEA2019	$\mathcal{S}$		$\mathcal{A}$	70.7	43.5	62.8	74.3	57.0	70.1
	Proposed+PIE-34K	$\mathcal{S}$		$\mathcal{S}$	73.9	39.5	**62.9**	78.1	53.5	71.5
	Proposed+C4-34K	$\mathcal{S}$		$\mathcal{S}$	75.1	37.5	62.5	79.7	52.2	**72.1**
	Original+BEA2019	$\mathcal{S}$	$\mathcal{A}$	$\mathcal{A}$	69.1	46.8	62.9	75.0	56.6	70.5
	PartiallyRecovered+BEA2019	$\mathcal{S}$	$\mathcal{A}$	$\mathcal{A}$	73.3	42.1	**63.9**	75.0	56.5	70.4
	Proposed+BEA2019	$\mathcal{S}$	$\mathcal{A}$	$\mathcal{A}$	73.4	41.5	63.6	75.8	55.3	**70.6**
C4-200M	Original	$\mathcal{S}$			64.2	39.1	56.9	62.9	50.4	59.9
	Proposed	$\mathcal{S}$			66.3	47.9	**61.6**	68.1	62.0	**66.8**
	Original+BEA2019	$\mathcal{S}$	$\mathcal{A}$		65.6	46.3	**60.6**	61.2	60.5	**61.0**
	Proposed+BEA2019	$\mathcal{S}$	$\mathcal{A}$		63.7	45.9	59.1	59.6	59.5	59.5
	Original+BEA2019	$\mathcal{S}$		$\mathcal{A}$	72.5	42.1	63.3	78.1	56.3	**72.5**
	Proposed+BEA2019	$\mathcal{S}$		$\mathcal{A}$	70.9	44.7	63.4	73.1	58.8	69.7
	Proposed+C4-34K	$\mathcal{S}$		$\mathcal{S}$	75.3	40.0	**64.0**	77.9	54.6	71.8
	Proposed+PIE-34K	$\mathcal{S}$		$\mathcal{S}$	74.8	39.9	63.6	78.2	54.3	71.8
	Original+BEA2019	$\mathcal{S}$	$\mathcal{A}$	$\mathcal{A}$	72.9	43.1	**64.0**	75.8	58.3	**71.5**
	Proposed+BEA2019	$\mathcal{S}$	$\mathcal{A}$	$\mathcal{A}$	73.4	41.4	63.6	75.1	56.3	70.4

提案した人工合成訓練データの再構成法は、GEC モデルの訓練データだけではなく、文の対からなるデータで学習するモデルに対して、広く応用することができる。

つぎに、再構成された人工合成訓練データの文法誤りの種類に関する統計的性質を明らかにした。文法誤りの種類は、図１に示す 24 種類とし、人工合成訓練データと再構成された人工合成訓練データそれぞれについて、データに含まれる文法誤りの種類に関する出現頻度分布を L2 学習者コーパスのうち、比較的品質がよい Cambridge English Write & Improve と LOCNESS(W&I+LOCNESS)および CoNLL2014 Shared Task の訓練データを比較した。ただし、

LOCNESS は、母語話者のエッセイのコーパスである。比較方法は、カルバック・ライブラー情報量 D_{KL} を用い、その結果を表２に示す。実験結果より、品質の比較的悪い人工合成訓練データ PIE に対して、訓練データを再構成すると、L2 学習者コーパスに統計的により類似するデータを得ることができることがわかった。

(2)意味構造を学習する RoBERTa モデルから構成されるサブモジュールを(1)の統語構造を学習する GEC モデルに統合して、2 つの独立した学習モデルが、統語構造と意味構造をそれぞれ学習し、統語構造と意味構造に関わるベクトルをニューラルネットワークの相互結合である最上位層で統合する GEC モデルを提案した。AMR パーサにより、文法誤りを含む文を AMR の木構造に変換し、さらに、 AMR 木構造データを先行順で操作して１次元トークン列とし、サブモジュールに入力した。AMR は、抽象概念であるため、上位概念で表現できること、抽象概念変換時に時制や前置詞の誤りを吸収することができると期待できることにより、文法誤り訂正能力が向上できると予想したが、実験より、統合前の GEC と同程度の訂正能力しか発揮できなかった。

(3)提案した GEC モデルを主要部とする文書作成支援アプリケーションソフトウエアを作成した。本ソフトウエアは、L2 学習者等が英文を入力すると、提案した GEC モデルにより文法誤りを指摘する機能を有する。同時に、Wordnet データベースから単語の概念と同義語を表示する機能を有する。図２は、本ソフトエアの画面の一例であり、文法誤りを含む英文を入力した例であり、2 つの文法誤りを指摘し、動詞 employ に関する 2 つの概念と 3 カテゴリ中の 7 つの同義語を表示している。本ソフトウエアは、第二言語の学習支援に寄与するものである。

図２　ソフトウエアの画面例

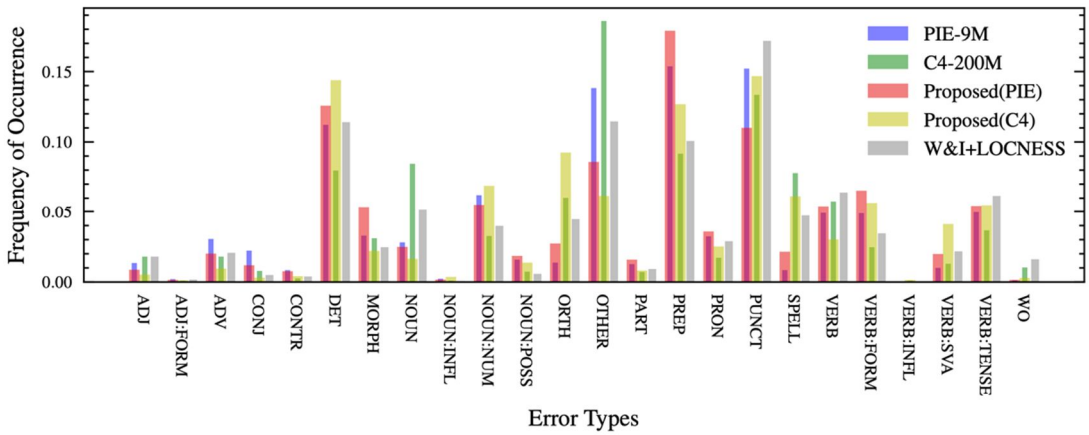

図１　人工合成訓練データの誤り種類に関する出現頻度分布の比較

表２　再構成された人工合成訓練データと L2 学習者コーパスとの文法誤り種類の発生確率比較

Synthetic (#sentences)	Dataset	$\mathbf{x}^i/\mathbf{y}^i/D$	#tokens	#edits	Entropy [bit]	$D_{\mathrm{KL}}(\mathcal{D}_{\mathrm{WI}}\|\|\cdot)$	$D_{\mathrm{KL}}(\mathcal{D}_{\mathrm{Co}}\|\|\cdot)$
PIE-9M		$\mathbf{x}^i$	25.1	—	—	—	—
(8.42M)		$\check{\mathbf{x}}^i$	25.2	—	—	—	—
		$\mathbf{y}^i$	25.4	—	—	—	—
		$\hat{\mathbf{y}}^i$	25.1	—	—	—	—
	Original	$\mathcal{D}(\mathbf{x}^i, \mathbf{y}^i)$	—	2.45	3.79	0.216	**0.198**
	Proposed	$\tilde{\mathcal{D}}(\mathbf{x}^i, \tilde{\mathbf{y}}^i)$	—	1.60	3.87	**0.186**	0.216
	Random	$\check{\mathcal{D}}(\check{\mathbf{x}}^i, \mathbf{y}^i)$	—	1.62	3.79	0.198	0.216
C4-200M		$\mathbf{x}^i$	25.7	—	—	—	—
(8.42M)		$\mathbf{y}^i$	25.7	—	—	—	—
		$\hat{\mathbf{y}}^i$	25.8	—	—	—	—
	Original	$\mathcal{D}(\mathbf{x}^i, \mathbf{y}^i)$	—	4.04	3.80	**0.093**	**0.177**
	Proposed	$\tilde{\mathcal{D}}(\mathbf{x}^i, \tilde{\mathbf{y}}^i)$	—	1.26	3.80	0.196	0.369
W&I+LOC		$\mathcal{D}_{\mathrm{WI}}(\mathbf{x}^i, \mathbf{y}^i)$	—	—	3.88	—	0.128
CoNLL2014		$\mathcal{D}_{\mathrm{Co}}(\mathbf{x}^i, \mathbf{y}^i)$	—	—	3.86	0.143	—

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Oda Mikio	４．巻 -
２．論文標題 Training for Grammatical Error Correction Without Human-Annotated L2 Learners’ Corpora	５．発行年 2023年
３．雑誌名 Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)	６．最初と最後の頁 455-465
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2023.bea-1.38	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0件

〔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