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 ２３ 日現在

機関番号：５７１０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ００９０４

研究課題名（和文）英語の語形変化推定を利用した言語モデルによる自動文書作成ソフトウエアに関する研究

研究課題名（英文）An automatic English sentence generator of a language model predicting the inflection of words

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）　　2,600,000円

研究成果の概要（和文）：工学系の学生に求められる技術論文の作成を支援するために，Webベースの支援ソフトウエアを開発した．支援ソフトウエアは，自然言語処理に関する論文から文数1.36M(データ量143MB)及び人工データを追加した文数22.3M(データ量1.49GB)のコーパスを用いて学習した言語モデルが主要部となっており，作文の先頭から数単語入力すると，引き続きの文を提案するものである．
本ソフトウエアは，工学系の学生が技術文書を作成するときにが有益に利用できると考える.また，本ソフトウエアに新しい機能を今後追加することにより，英語能力が初級の学生に対する高機能な支援に発展させることができると考える．

研究成果の学術的意義や社会的意義
近年，インターネット上のビッグデータを様々な応用問題に活用する試みがされている．本研究は，ビッグデータであるコーパスを用いて構築される言語モデルを工学系の学生が英作文をするときに役立てることを目的とした．本研究において，最新の言語モデルの改良や評価検討には，学術的な意義がある．また，自然言語処理に関する論文から大規模なコーパスを作成して学習した言語モデルを工学系学生に活用してもらうためのWebベースのソフトウエア開発は，社会的な意義があり，今後，本システムをより高機能化することにより，第二言語習得の教育に寄与することが見込まれる．

研究成果の概要（英文）：We developed web application software which support students in science research writing. The software is mainly composed of a language model and the model is trained with text corpora of 1.36M sentences (143MB) and 22.3M sentences (1.49GB) generated from technical papers on natural language processing. When the beginning part of a sentence is inputed, the software generates the following part of the sentence.
We believe that the software is very useful and helpful when students work on writing their own research papers. Also, the software can extend the functions of rendering other linguistic knowledge to students of various skill levels.

研究分野：自然言語処理

キーワード：言語モデル　第二言語習得　英文作成支援　文法誤り訂正

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
(1)言語の確率モデルは、文法に基づく演繹的な方法論をとる理論言語学、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る大量データを用いた応用的成功により、単語の表現学習および言語モデルの学習に対して、大きな技術的発展の可能性を秘めている。

(2)計算言語学の学術的成果の応用例として、音声の自動翻訳等があるが、教育支援関係については、英文技術文書の作成支援への応用が考えられる。近年の目覚ましいグローバル化に伴い、学生や技術者が英文技術文書を作成する要請が高まっており、外国語の文書を作成するときに、続きの文の候補を提示するアプリケーションの開発が望まれる。

２．研究の目的
(1)単語は語形変化があり、複数の品詞等の役割をする単語の多様性に関する検討をするため、確率的言語モデルを用いて、単語の基本形から語形変化を生成する手法を検討し、正しい品詞や語形変化の単語をどれくらいの精度で文脈から推定できるかを明らかにする。

(2)英文の技術文書を作成支援するソフトウエアは、日本人学生や技術者にとって非常に有益であるため、ユーザが入力した一部の文の文脈から次単語等の候補を提示する Web アプリケーションを開発することを目的とする。

３．研究の方法
(1)タグに基づき単語の基本形から変化形を生成する機構は、CoNLL2017 Shared Task で検討されており、sequence-to-sequence モデルによる学習が最良の結果を得ている。Shared Task では、単語のアルファベットのみを用いて学習しているが、提案手法により、動詞の変化形推定を sequence-to-sequence モデルで推定し、提案手法の推定精度を評価する。

(2)工学系の学生に求められる技術論文の作成を支援するためには、作文の先頭から数単語入力すると引き続きの文を提示する言語モデル等の単語推定モデルが必要である。広い意味での言語モデルには、sequence-to-sequence モデル、分類モデル、マスク言語モデルがあり、どのモデルを採用するかを実験することにより決定する。また、採用した言語モデルを効果的に学習するために、技術分野を自然言語処理に限定して、自然言語処理に関する大規模な技術文献のコーパスを作成する。

(3)英文技術文書作成支援ソフトウエアのインターフェースを開発する。http プロトコルによる遠隔使用も可能とし、多くの学生が利用できるように、Web ベースのソフトウエアとする。言語モデルを Python 言語のライブラリでコーディングしているため、インターフェースも Python 言語および Flask ライブラリを用いて開発する。

４．研究成果
(1)動詞等の変化形推定は、発音に関連するとの考察から、変化形の推定精度を向上させるために、単語のアルファベットだけでなく、発音記号を入力および出力に用いることを提案した。言語モデルの学習モデルの一つである Long short-term memory(LSTM)による sequence-to-sequence モデルを用いて、アルファベットと発音記号のどちらが語形変化の推定学習に有効であるかを数値実験により検証した。実験の結果、発音記号を用いた学習法より、アルファベットを用いた従来法に優位性があり、提案手法は、CoNLL2017 Shared Task の最良結果を超える精度を得ることはできなかった。

(2)POS Tagger を用いて British National Corpus を単語の基本形と品詞の組にし、sequence-to-sequcence モデルを学習し、言語モデルを構築した。British National Corpus を直接学習した言語モデルと提案言語モデルについて、数値実験により次単語の推定能力を比較すると、British National Corpus を直接学習した言語モデルが、より能力が高いことがわかった。(3)上記の研究成果を踏まえ、言語モデルでは、単語の基本形を採用せず、単語の変化形を含む Wikipedia のコーパスを用いて、sequence-to-sequence モデルを学習した。sequence-to-sequence モデルとして、Bidirectional Encoder Representations from Transformers (BERT)を用いた。ここで、言語モデルによる英文技術文書作成支援ソフトウエアへの応用を前提に、日本人がよく間違う前置詞推定の能力を実験で確認した。予備実験として、L2 学習者コーパスである the Cambridge English Write & Improve(W&I)のデータを分析し、L2 学習者が、誤用す

る前置詞の組を調べた。統計分析結果を示す図 1 のとおり、前置詞の in と of を誤用する傾向があることがわかった。さらに、Wikipedia で学習した BERT 言語モデルを用いて、L2 学習者の文法誤りを含むコーパス W&I の文を誤り訂正した。図 2 に示すとおり、L2 学習者と同様、言語モデルによる誤り訂正においても、前置詞の in と of を誤推定しやすいことがわかった。

(4)前置詞の誤り訂正モデルとして、 wikipedia を事前学習し、L2 学習者コーパスを事後学習とした分類モデルを提案した。提案した分類モデルは、上記の実験結果に基づき、頻出の前置詞をグループ化し、BERT の入出力を多段に接続した推論モデルとした。具体的には、第 1 分類段階で、前置詞の in と of とその他を分類し、第 2 分類段階で、その他の前置詞を分類するモデルとした。実験結果より、提案する分類モデルは、従来の言語モデルより、全体の推定精度は、若干低下したが、注目した前置詞 in と of の推定に関して、適合率が若干改善した。

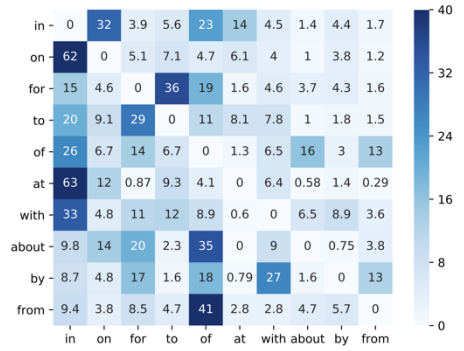

図 1 正解前置詞と誤り前置詞の関係

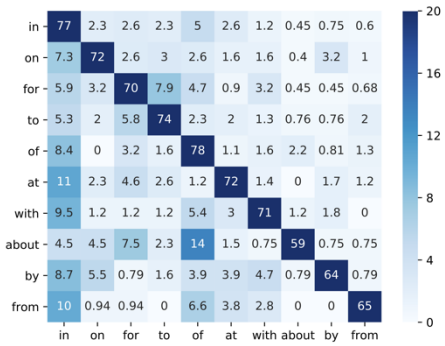

図 2 言語モデルによる正解前置詞と誤り前置詞

表 1　提案分類モデルの前置詞誤り訂正能力

前置詞	訓練データ中の比率 [%]	予備実験 R	予備実験 P	予備実験 $F_{0.5}$	$M_3 + M_m$(in, of, [OTHERS]) R	$M_3 + M_m$(in, of, [OTHERS]) P	$M_3 + M_m$(in, of, [OTHERS]) $F_{0.5}$	$M_4 + M_m$(in, of, to, [OTHERS]) R	$M_4 + M_m$(in, of, to, [OTHERS]) P	$M_4 + M_m$(in, of, to, [OTHERS]) $F_{0.5}$
in	**16.9**	0.771	0.679	0.695	0.719	**0.700**	0.703	0.731	**0.696**	0.703
on	5.6	0.715	0.810	0.789	0.725	0.774	0.764	0.727	0.789	0.776
for	7.6	0.703	0.726	0.721	0.703	0.703	0.703	0.703	0.703	0.703
to	**8.9**	0.738	0.738	0.738	0.743	0.728	0.731	0.720	**0.735**	0.732
of	**27.2**	0.776	0.667	0.686	0.774	**0.687**	0.702	0.757	**0.690**	0.703
at	4.6	0.722	0.798	0.782	0.722	0.793	0.778	0.722	0.801	0.784
with	5.8	0.714	0.600	0.620	0.726	0.589	0.612	0.726	0.589	0.612
about	1.2	0.586	0.743	0.705	0.602	0.690	0.670	0.617	0.633	0.669
by	4.6	0.638	0.618	0.622	0.591	0.615	0.610	0.575	0.635	0.622
from	3.8	0.651	0.616	0.623	0.660	0.560	0.578	0.642	0.553	0.569
TOTAL		$AC = 0.691$			$AC = 0.680$			$AC = 0.680$		

(5) 工学系の学生に求められる技術論文の作成を効果的に支援するためには、用いる言語モデルの学習コーパスは、Wikipedia などの日常文より、技術論文がより望ましい。また、技術分野を特定するほうが、効果的な支援ができると考え、技術分野を自然言語処理とし、Association for Computational Linguistics の論文誌 33 誌の 2000 年から 2021 年までの論文を収集した。収集した PDF ファイルからテキストを抽出してクリーニングし、さらに、アルファベット大文字から始まりピリオドで終了する文のみに限定した。その結果、作成したコーパスは、1.36M の文数で、テキストデータの容量は、143MB となった。さらに、使用する言語モデルは、双方向アテンションを利用するマスク言語モデルであるが、一方、支援ソフトウエアが提示する単語は、前方向の単語のみから推定する必要があるため、コーパスの各文について、先頭から任意の単語までの部分文章を人工生成したコーパスも作成した．このコーパスは、22.3M の文数で、テキストデータの容量は、1.49GB となった。

(6)上記の検討結果を踏まえ、マスク言語モデルを英文技術文書支援ソフトウエアの主要部に採用することとした。マスク言語モデルである BERT, BART, RoBERTa 等の言語生成能力を比較し、十分な能力を有する BERT を採用した。BERT を上記作成したコーパスで学習したプログラムとインターフェースのプログラムとを連携することができた。インターフェースのプログラムは、Python 言語と Flask ライブラリを用い、そのインターフェース画面例を図 3 に示す。図 3 のインターフェース例は、ユーザが入力した英文の始めの部分から次の単語の候補および引き続き

の文を提案する機能を示したものである。本システムは、工学系の学生が技術文書を作成するときに有益に利用できると考える. また、本システムは、新しい機能を今後も追加して、英語能力が初級の学生に対する高機能な支援に発展させることができると考える.

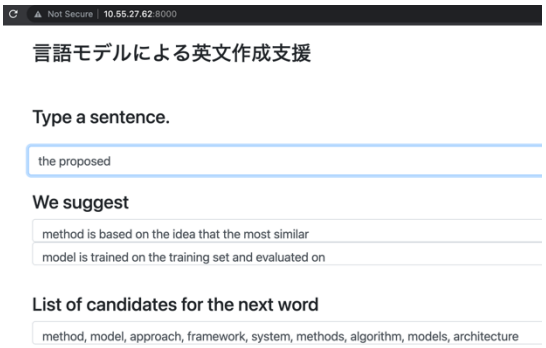

図 3 支援ソフトウエアのインターフェース例

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 小田幹雄
２．発表標題 分類器による英文前置詞誤り訂正の学習法
３．学会等名 言語処理学会 第26回年次大会 発表論文集
４．発表年 2020年

〔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