氏　　名　壹岐 太一

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2364 号

学位授与の日付　2022 年 9 月 28 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Language Models for Task Execution via Graphical User Interfaces

論 文 審 査 委 員　主　査　相澤 彰子
情報学専攻 教授
高須 淳宏
情報学専攻 教授
杉本 晃宏
情報学専攻 教授
稲邑 哲也
情報学専攻 准教授
菅原 朔
情報学専攻 助教
河原 大輔
早稲田大学 基幹理工学部 教授

（様式３）

博士論文の要旨

氏　　名　壹岐　太一

論文題目　Language Models for Task Execution via Graphical User Interfaces

Language models, the term used in this thesis to refer to pre-trained transformer approaches, use large amounts of data (e.g., text) to train the transformer structure to recognize patterns in the data and then further tune the structure for a specific task. After their initial success with text data, language models have been extended to perceptual data, such as images and audio, as well as to the fusion of heterogeneous data. With the advent of language models, the integration of linguistic and perceptual capabilities of artificial intelligence is rapidly advancing.

In this thesis, we argue that this integration should be extended to the modality of action generation. We expect this to benefit language models' linguistic and behavioral capabilities. Integrating these capabilities within a single model will enable the model to understand action statements in natural language texts more deeply. Furthermore, understanding actions through language makes the model's ability to generate action more flexible.

Previous studies have applied language models to action generation in simulated three-dimensional (3D) spaces. These tasks require the models to follow instructions by taking action (e.g., moving forward, rotating, and grasping objects). However, these studies have not covered the space of computers with which we interact via graphical user interfaces (GUI). For simplicity, we consider desktop GUIs. There are several differences between simulated 3D space and GUI space: (1) system of actions: we control GUIs by combining actions, such as moving a cursor, clicking, and pressing keys, (2) visual elements: the screen comprises icons, windows, and buttons, and (3) use of text: the text is rarely displayed in 3D space whereas it is the main content in the GUI space. These differences necessitate the exploration of language models for the GUI space.

To open the field of GUI control with language models, we created a task set with GUIs and proposed language models that perform tasks by controlling the user interfaces. To create our model, we added screen image input and a memory mechanism to BERT, a language model trained on text data, and trained it on action selection.

This thesis consists of seven chapters. In Chapter 1, we describe the motivation and discuss the task scope of our model. We begin with Nilsson's employment test, which measures a piece of artificial intelligence by its ability to perform human jobs. Inspired by this test, we set a long-term goal for the model to perform annotation tasks like humans using GUI, a task commonly performed in crowdsourcing. We analyzed the GUI

templates for these tasks and identified the requirements.

In Chapter 2, we provide a technical background. We describe the transformer structure, language models, and their visual and linguistic fusion applications.

In Chapter 3, we investigate how additional visual training for language models affects their ability to solve text tasks, that is, tasks intended to be solved only with text. As the text is the primary content in GUIs, we need to understand such an impact. To estimate the impact, we focused on five vision-and-language BERTs enabled to accept visual inputs by additional training with image-caption pairs. We evaluated them on nine text tasks. We observed that all models scored lower than the original language model, but the difference was slight. This result suggests that additional training does not significantly break the linguistic ability of the model, although it is not fully maintained.

In Chapter 4, we develop a task set with a GUI and create our models called BUI-BERT. First, we created the task set by adding GUIs to existing benchmark tasks in natural language processing and vision-and-language domains. Second, we created BUI-BERT by further training the BERT on actions. Finally, we fine-tuned BUI-BERT to benchmark tasks with a GUI for evaluation. We observed that BUI-BERT learns to submit responses using the GUI but is less accurate than BERT using the original task format. These results suggest room for improving the BUI-BERT task performance. The ablation analysis shows that our additional training and memory mechanisms were effective.

In Chapter 5, we focus on the visual input. BUI-BERT uses a pre-trained image recognition model to convert a screen image into a feature map. This model is independent of the text input. To fuse early-stage visual processing and language input, we propose a method for creating a text-aware feature map by modulating the representations inside the model with text. We evaluated our method on two datasets of referring expression comprehension. The results show that our method performs as well as or better than existing methods. The ablation analysis confirmed the effectiveness of this method. Incorporating this method into BUI-BERT is expected to improve the performance of vision-related tasks.

In Section 6, we describe the prospects of the study. We summarize the remaining challenges in solving the annotation task using GUIs. We also discuss task unification. In Section 7, we summarize the contributions and conclusions of this study.

In summary, we proposed a GUI control using language models. This study complements the study of action generation in a 3D space. We expect the GUI approach to evolve into a flexible automation technique by overcoming the challenges identified in this study.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　壹岐　太一

Title
論文題目　　Language Models for Task Execution via Graphical User Interfaces

出願者は、グラフィカルユーザーインターフェース（GUI）操作を介して課題を解く環境を設計し、学習済み言語モデルを拡張した GUI 操作モデルを構築して、実験において有用性を実証した。これにより、環境の中で入出力を伴う行動を自動生成して、高度な自然言語処理を伴う多様な課題を解くことができるモデルの実現法を示した。

本学位論文は、「Language Models for Task Execution via Graphical User Interfaces（グラフィカルユーザーインターフェースを介したタスク実行のための言語モデル）」と題し、全７章から構成されている。

第１章では、研究背景と提案手法の概要を説明している。はじめに、人と同じように問題解決する人工知能の実現には環境の中で行動系列を自動生成して課題を解くモデルの研究が必要であると問題提起し、言語や画像の理解モデルと行動生成を組み合わせる研究の不足を指摘している。次に、既存の行動生成の研究環境として３次元空間上でのシミュレーション世界があるが、そこではテキストを読む行動が対象とならないため、環境の中で問題を解くモデルの研究には不十分であると論じている。以上の背景のもと、コンピュータ GUI の操作に関する行動を生成して課題を遂行するモデルの研究を提案した上で、その研究の長期的な目標としてブラウザユーザーインターフェース（BUI）を用いたアノテーションタスクを導入し、要件を整理している。最後に、学習済み言語モデルを拡張する提案手法の概要を説明している。

第２章では、まず言語モデルと視覚言語モデルの歴史を概観し、記号接地問題の観点からの行動生成の研究の重要性を議論している。続いて、前提知識として、言語モデルの構造的な基礎である Transformer 構造や視覚言語モデルの技術的な説明を述べている。最後に関連分野の研究として、身体性を有するエージェント、ユーザーインターフェースモデリング、マルチタスク、文書画像の理解の分野における言語モデルの応用を取り上げ、本研究との関係を論じている。

第３章では、言語モデルの追加学習による視覚への拡張が、テキストのみで解くことを意図したタスク（以下、テキストタスク）の性能に与える影響を評価している。画像キャプション対で追加学習した５つのモデルについて９つのテキストタスクにおける性能を評価し、どのモデルも元の言語モデルに比べてテキストタスクの性能が低下するが、その差は小さいことを明らかにしている。この知見から、モデルの言語能力は追加学習では大きく損なわれず、追加学習の手法を GUI 操作への拡張にも適用できると結論づけている。

第４章では、まず BUI 操作環境と操作モデルを作成し、続いてモデルの評価を行っている。BUI 操作環境の作成では、マウスやキーボードを用いた汎用的な行動を定義し、行動

に関する事前学習タスクをルールにより自動生成している。また、言語や画像に関する既存データセットに BUI を付与してベンチマークタスクを作成している。操作モデルの作成では、言語モデルにスクリーンショットを扱うための視覚入力と記憶構造を追加して、行動に関する事前学習タスクで行動生成の追加学習をすることでモデルを作成している。モデルをベンチマークタスクで学習・評価し、操作モデルは BUI 操作を学習できることと、正解率のさらなる向上のためには言語モデル自体の性能向上が必要であることを結論づけている。また、アブレーション分析の結果から、記憶機構と行動に関する事前学習タスクの有効性を確認している。

第 5 章では、視覚モデルの改良について論じている。第 4 章の操作モデルには視覚情報の処理において言語情報を考慮することができないという制約がある。その制約の解消に向けて、言語による条件付けで視覚モデルの内部特徴量を再構成する手法を提案している。参照表現理解に関する 2 つの既存データセットを用いた評価によって、同一のデータを用いて学習した既存手法に対して提案手法は同等以上の性能が得られることを実験的に明らかにしている。また、アブレーション分析によって、言語による条件付けが性能の向上に有効であることを示している。

第 6 章では、言語モデルを用いた GUI 操作の課題と展望を述べている。まず、個別タスク性能、汎化性能、構造的制約の 3 つの観点から性能改善に向けた課題を考察している。また展望として言語モデルを用いた GUI 操作によるタスク統一の構想について議論している。

第 7 章において、各章の貢献をまとめ、結論を述べている。

以上を要するに、本学位論文で提案された環境は汎用的な行動でブラウザを操作するモデルの訓練と評価を可能にし、作成されたモデルは今後のモデル設計に対して有用な知見を提供するものである。本学位論文は、学習済み言語モデルの応用方法の追求という重要な課題に対して、マウスやキーボードを用いた汎用的な GUI 操作という新しい方針を提示するものである。

公開発表会では博士論文の章立てに沿って発表が行われた。その後に行われた論文審査会および口述試験では、審査員からの質疑に対して適切に回答がなされた。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究は十分なクオリティとオリジナリティがあるとの評価がなされた。また、本学位論文の成果は、情報学専攻が定めるトップ国際会議のフルペーパー(Findings カテゴリ) 1 件、査読付きショートペーパー国際会議論文 1 件として発表され、学術的な貢献も認められる。以上の理由により、審査委員会は本学位論文が学位の授与に値すると判断した。