

UGIF: UI に基づいた指示に従う

サーガル・グब्ビ・ベンカテッシュ

Google リサーチ
インド

gubbi@google.com

パーサ・タルクダル

Google リサーチ
インド

partha@google.com

スリーニ・ナラヤナン

Google リサーチ
スイス

srinin@google.com

概要

新しいスマートフォンユーザーは、スマートフォンを操作するのが難しく、通話やメッセージなどの限られた機能しか使用しないことがよくあります。これらのユーザーは、スマートフォンを使って探索することを躊躇し、経験豊富なユーザーに電話の使い方を教えてもらいます。ただし、経験豊富なユーザーが常にガイドをしてくれるとは限りません。新しいユーザーが電話を自分で使用する方法を学習できるように、UI を介して操作し、さまざまなタスクの実行方法をユーザーに示すエージェントに従う自然言語ベースの指示を提案します。「不明な番号からの着信をブロックするにはどうすればよいですか?」などのよくある質問は、サポートサイトに記載されており、ユーザーが何をすべきかを自然言語で説明した一連の手順が記載されています。

大規模言語モデル (LLM) を使用してこれらのステップを解析し、ユーザーがクエリを要求したときにデバイス上で実行できるマクロを生成します。このエージェントを評価するために、スマートフォンで段階的にタスクを完了するための多言語、マルチモーダルUIデータを導入します。523 の自然言語命令と、多言語 UI 画面のペアシーケンスと、8 つの言語でタスクを実行する方法を示すアクションが含まれています。PaLM、GPT3 などのさまざまな大規模言語モデルのパフォーマンスを比較したところ、エンド ツー エンドのタスク完了成功率は、英語の UI では 48% ですが、英語以外の言語ではパフォーマンスが 32% に低下することがわかりました。このタスクに関する既存のモデルの一般的な故障モードを分析し、改善すべき領域を指摘します。

1 はじめに

スマートフォンを初めて使用するユーザーは、そのユーザー インターフェイスに問題を抱えています。この問題は、読み書き能力のレベルの違いや、電話の所有コストが高いなどの理由で、発展途上国で特に深刻です。彼らは、電話のインターフェイスを試すことを躊躇し、通話やメッセージなどの基本的な機能しか使用しないことがよくあります。

¹ソフトグで発音 :U-JIF

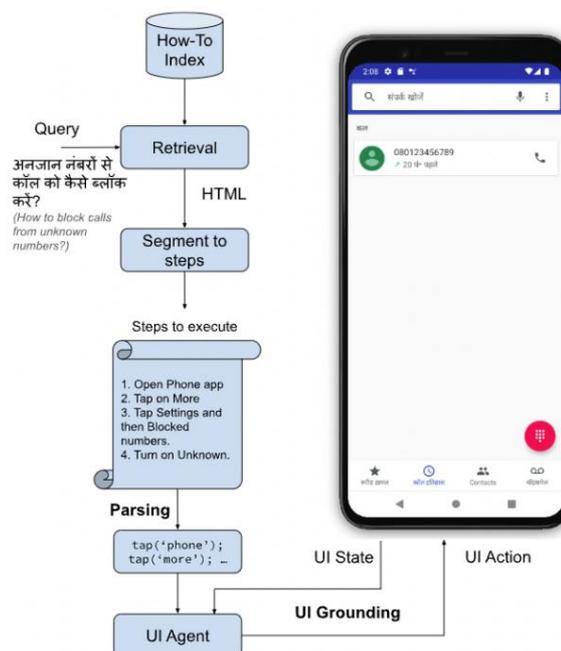


図 1: 私たちのモデルは、ハウツー ステップを解析し、UI にグラウンディングすることでデバイス上で実行できる、tap()、toggle()、home() などのマクロを生成します (セクション1)。

したがって、彼らはスマートフォンとインターネットが提供する価値を十分に活用することができません(Ranjan, 2022)。新しいユーザーがデジタルの世界に簡単に参加できるようにするツールの 1 つは、音声対話です。音声コマンドは、電話との対話を簡単に開始する方法を提供しますが、プライバシーや社会的理由から、ユーザーは常に音声コマンドを使用したいとは限りません。そのため、音声インタラクションは、GUI の代替としてだけでなく、音声クエリに応じてUI をナビゲートする方法をユーザーに示すことで、GUI についてユーザーに教えるオンボーディング デバイスとしても見なすことをお勧めします。

新規ユーザーからのよくある質問 (FAQ) の多くは、スマートフォン自体に関するものです。これらのクエリは、Wi-Fi や Bluetooth の状態、バッテリー セーバーなどのデバイス設定を変更する方法、電話番号をブロックする方法、または変更する方法に関するものです。

ブラウザのプライバシー設定。このような FAQ は、 <https://support.google.com>などのサポート Web サイトに記載されています。ハウツー クエリのサポート ドキュメントには、タスクを完了するために実行する必要がある一連のアクションをユーザーに知らせる、自然言語による段階的な手順が含まれています。ユーザーが仮想エージェントがどのようにタスクを実行するかを確認し、デモンストレーションを観察することで UI の使用に自信を持てるように、これらのハウツードキュメントを電話の UI で実行できるようにすることを検討します。私たちの目標は、新しいユーザーが電話に参加できるようにすることですが、この支援技術は、ユーザーが料理中や運転中に電話の画面に触れられないなど、状況に応じて障害が発生している他のユーザーにも役立ちます(Sarsenbayeva, 2018)。

私たちのアプローチは、タスクを 2 つのコンポーネントに分割します。解析ステップとグラウンディング ステップです。サポート サイトがクロールされ、検出されたハウツー手順の手順が大規模な言語モデル(Chowdhery et al., 2022)を微調整して解析され、tap()、toggle()、home()などのマクロが生成されます。一致するクエリがユーザーによって発声されると、対応するマクロ シーケンスが、UI 内の各マクロをグラウンディングすることによって実行されます。このために、多言語文埋め込みモデル(Feng et al., 2020)を使用して、最も一致する UI 要素を見つけます。

メント。

UGIF-DataSetと呼ばれる新しい多言語、マルチモーダル UI に基づいたデータセットを収集して、Android UI でハウツー命令をどれだけうまく実行できるかを評価します。これは 523 のハウツー クエリで構成され、クエリごとに、英語での指示手順と、ハウツーを完了するための一連の UI スクリーンショットとアクションで構成されています。各ハウツー クエリと UIシーケンスは 8 言語で利用できます。このデータセットの構造の概要を図2 に示します。

この作品の貢献は次のとおりです。

- UGIF-DataSetをリリースします。これは、ヒューマン アノテーターによって実行されるハウツー クエリと一連の UI 画面およびアクションの新しい多言語、マルチモーダル データセットです。これは、この種の最初のマルチモーダル データセットです。
- 大規模な言語モデルを使用した手順での段階的なハウツーの解析と、多言語 BERT 文の埋め込み (LaBSE)を使用した UI接地を評価します。
- 私たちの結果は、特に英語以外の言語で、パフォーマンスを改善する余地がかなりあることを示しています。さらに、

アプリの設計が時間の経過とともに進化するにつれてバージョンが変更されることによる UI の不一致は、重大なエラーの原因であり、研究とエンジニアリングの両方の課題を提示します。

2 関連作品

UI ナビゲーションに続く自然言語命令:デスクトップ オペレーティング システム(Branavan et al., 2009, 2010; Xu et al., 2021)および Adobe Photoshop などの画像編集アプリケーション向けの自然言語調整された UI ナビゲーションには、以前からいくつかの取り組みがありました。(マヌピナクリケラ、2018)。最近では、ヘルプ記事のビデオを自動的に生成するためのモバイル ユーザー インターフェイスへの指示に自然言語を使用する作業が行われています(Zhong et al., 2021)。私たちの仕事は、Li らでリリースされたPixelHelpデータセットの拡張および更新された後継です。(2020a)は、8 か国語での音声およびテキスト クエリ、英語での指示ステップ、および8 つのシステム言語での UI 画面を備えています。

UI ナビゲーションのための模倣学習と強化学習: UI ナビゲーション エージェントを構築するには、大まかに次の 2 つのアプローチを考えることができます。これは多くの異なるアプリで有用であり、(b)いくつかのアプリケーションに対してのみ、より深い機能を公開することで垂直方向にスケールします。Li (2021)は former アプローチを採用し、振る舞いの複製と強化学習を使用してエージェントに 2 つの特定のスキルをトレーニングします: Playストアから指定されたアプリをインストールするスキルと、ユーザーが任意のアプリケーションで何を求めているかを最初に見つけることによって検索する別のエージェント検索ボックス。Android UI で強化学習の研究を可能にするために、Toyama et al. (2021)は、RL エージェントをトレーニングするためのオープンソースプラットフォームであるAndroidEnv を紹介します。それと同様に、WorldOfBitsはWeb ナビゲーション エージェントをトレーニングするためのオープン プラットフォームです(Shi et al., 2017; Liu et al., 2018)。私たちの仕事では、Android サポート サイトのヘルプ記事を利用して、いくつかの人気のあるアプリのより深い機能を公開するという後者のアプローチを採用しています。これを選択したのは、新規ユーザーが、特定のアプリの操作方法に関するより深い知識を必要とする目的指向の質問をすることが多いためです。さらに、アプリ開発者は一般的なタスクを念頭に置いて FAQ を提供することが多いため、サポート ページを活用して、新しいユーザー向けの UI に基づいたチュートリアルを作成できます。

UI タスクの事前トレーニング:過去数年間で、事前トレーニングと微調整へのディープラーニングのパラダイムシフトがありました。Foundationモデルは、ラベル付けされていない広範なデータセットで、自己教師ありの学習目標を使用して事前にトレーニングされています。このような事前トレーニング済みモデルの微調整は、小さなデータセットでの教師あり学習のみと比較して劇的な利点をもたらしました。パイラ。(2021);彼等。(2021) UIにこのアプローチを採用し、Web クローリングと同様の方法でスマートフォンのアプリをクロールして取得した多数のスクリーンショットでトランスフォーマーモデルを事前トレーニングします。これらの事前トレーニングされたUIBertおよびActionBERTモデルはリリース済みです。

UIグラウンディングタスクのための作業を行います。私たちは多言語UI画面に重点を置いているため、UIの基礎として事前トレーニング済みのLaBSE (Feng et al., 2020)を使用することを選択しましたが、広範なUIデータを利用することは、将来の改善のために重要です。

大規模な言語モデル:ウェブからスクレイピングされた大規模なテキストコーパスで事前トレーニングされた大規模な言語モデル(LLM)は、驚くべき少数の一般化機能を示しています(Chowdhery et al., 2022; Brown et al., 2020)。ヘルプ記事の解析にLLMを採用しています。ただし、プライバシー上の理由からデバイス上でUIグラウンディングを実行し、劣悪なネットワーク条件に対して堅牢であるため、UIグラウンディングにLLMを採用していません。

人間とロボットの相互作用における言語の基礎:人間とロボットの相互作用のための言語ガイド付きロボットアクション(Lynch and Sermanet, 2020; Venkatesh et al., 2021)は、自然言語駆動型のUIナビゲーションの問題に広く関連しています。

ただし、UIは、画面上で観察されるものと実行できるアクションの両方で構造的に離散的ですが、ロボットの観察とアクションはどちらも連続的です。ロボットでは、ビジョンモデルを使用してカメラフィードからシーングラフを推測する必要がありますが、ユーザーインターフェイスでは、多くの場合、画面のビュー階層を直接利用できるため、生のピクセルを処理する必要はありません。同様に、実際のロボットでアクションを実行することは、はるかに複雑で結果が不確実ですが、UIでは正確なアクションをほぼ確実に実行できます。その結果、UIに基づいたインタラクションの難しさは、センシングやアクチュエーションではなく、ユーザーの意図を理解し、サポートページなどの外部リソースを使用してアプリの構造を理解することでアプリをナビゲートすることです。

アイコンとウィジェットのキャプション: Android UI システムでは、開発者はコンを提供できますが、

```
PixelHelp++ dataset structure
[
  {
    "query": "How to block calls from unknown numbers?",
    "query_i18n": {
      "hi": ["...", "..."],
      ...
    },
    "query_i18n_speech": {
      "hi": ["...<wav_file>..."],
      ...
    }
  },
  "instruction_txt": "1. Open the Phone app...",
  "instruction_mark": "1. Open the 'Phone' app...",
  "macros": "tap('Phone');...",
  "url": "https://support.google.com/accessibility/...",
  "url_content": "<!DOCTYPE html>..."
  "ui_screens": [
    {
      "ui_xml": "<?xml ...",
      "ui_screenshot": "...<png_file>...",
      "ui_elements": [
        {
          "ui_str": "Phone",
          "ui_str_i18n": {
            "kn": "ಫೋನ್",
            "hi": "फोन",
            ...
          },
          "ui_bbox": [0, 0, 10, 20],
          "ui_type": 0, // button, switch, etc.
        },
        ...
      ],
      "ui_action": 0 // index into "ui_elements"
    },
    ...
  ]
},
...
]
```

図 2: UGIF-DataSet データセットの概要。523 組の操作手順と一連の UI 画面とアクションで構成されています (セクション3)。

すべてのアプリ開発者がそうしているわけではありません。幅広いアプリをサポートするには、アイコンとウィジェットを認識する必要があります(Li et al., 2020b; Baechler and Sunkara, 2021)。私たちの作業では、すべてのアプリが必要な説明を提供するため、アイコンのキャプションは必要ありません。

3 UGIF-DataSet: 新しい多言語マルチモーダル UI に基づいた指導次のデータセット

私たちの目標は、初心者ユーザーに Android UI でタスクを実行する方法を教えることができる UI ナビゲーション エージェントを構築することです。このようなエージェントを構築し、そのパフォーマンスを評価するために、UGIF-DataSet2と呼ばれる新しい多言語、マルチモーダル UI に基づくデータセットを収集します。ハウツークエリのコーパスです

[2https://github.com/google-research/google-research/tree/master/ugif](https://github.com/google-research/google-research/tree/master/ugif)

大きい	関数
テープ)	引数で指定された UI 要素をタップ(e)
トグル (e, val = True)	引数(e)で UI 要素を見つけてから、最も近い Switch 要素を検索し、それをタップします。
家 ()	アンでホームボタンを押す ドロイド
back()	戻るボタンを押します
prompt(a)ユーザー	にアクション(a)を実行するように要求し、アクションが実行されるまで待機します。

表 1: 命令ステップから生成できるすべてのマクロのリスト (セクション3)。

複数の言語によるテキストと音声で、各ハウターの説明ステップが一連の UI 画面とアクションと組み合わせられ、ハウターとしての操作が行われます。

さまざまな UI 言語設定の Android デバイスでヒューマン アノテーターによって作成されました(図2)。

Pixel ヘルプのサポート ページには、Androidで一般的なタスクを実行するための段階的な手順が記載されています。これはタスクの例です: 「不明な番号をブロックする方法は?」指示文が「1.電話アプリを開きます。2. [その他] をタップします。3. [設定],[ブロックする番号] の順にタップします。4. 「不明」をオンにします。Android サポート サイトをクロールし、ヘッダーの下の順序付きリストを探し単純なルールを使用してハウター手順を抽出します。ハウター ステップは、アノテーターによって表1の一連のマクロに解析されます。

ハウター タスクごとに、アノテーターは仮想 Android デバイスを操作してハウターの手順を実行し、デバイスの画面と

アノテーターのアクションが記録されます。アノテーターが実行した各アクションが仮想デバイスに転送され、UIAutomator (Android、2022 年)を使用して実行される直前に、デバイスのスクリーンショット、XML のビュー階層、およびそのステップでアノテーターが実行したアクションを記録します。アノテーターが各ステップで実行できるアクションを、(a) UI 要素をタップする、(b) ホーム ボタンを押す、(c) 戻るボタンを押す、(c) エンドユーザーに入力、(d) スイッチ/チェックボックスの切り替え、(e) 上下のスクロール、(f) タスクの完了の通知、(g)ハウター指示テキストのエラーの通知と完了前の記録の終了。

UI を収集するための手動アノテーション プロセス

Android エミュレーターの画面は、UI 言語の数に比例してスケールリングされます。これを軽減するために、アノテーターから英語のみの UI 画面を収集し、アプリの APKのリソース ディレクトリで各 UI 文字列を検索し、利用可能な場合はAPK で開発者が提供する翻訳に置き換えます。翻訳が利用できない場合は、デフォルトで英語になります。通常の UI 画面では英語とその他の言語の文字列が混在していますが、これは1 つの文に2 つの言語が混在するコードの混在とは異なります。

UGIF-DataSetデータセットには、152 (トレーニング) / 106 (開発) / 265 (テスト) のサンプルがあります。次のアプリのタスクが含まれます: 設定、Google One、Gmail、Play ストア、連絡先、メッセージ、Chrome、マップ、カメラ、Google フォト、Google Earth、およびファイル。UGIF-DataSetは、次の点でPixelHelp データセット(Li et al., 2020a)とは異なります。

- 英語以外の7つの言語(ヒンディー語、カンナダ語、マラーティー語、グジャラート語、ベンガル語、スワヒリ語、スペイン語)の UI 要素が含まれています。
- 「ごみ箱に移動したいメールを選択してください」などのユーザー入力が必要な操作手順が含まれています。
- 含まれていないマルチモーダル データセットです。画面のビュー階層だけでなく、実行の各ステップでのスクリーンショットも含まれます。
- UI 要素が可視であることを前提としない画面に表示されます。アノテーターは、スクロールして、説明テキストで参照されている UI 要素を見つけます。
- 説明テキストが古く、現在のバージョンの UI に対応していないサンプルが含まれています。このような場合、アノテーターは現在の UI に指示を適応させるか、タスクを完了できない場合はエラーを宣言することができます。

4 モデル

UGIF には、検索、解析、グラウンディングの3つのコンポーネントがあります。テキストまたは音声入力に基づいて、最も関連性の高い英語のハウター命令が取得され、解析されてマクロが生成されます。これらのマクロは、UI に接地することで Android デバイス上で実行されます (Alg. 1)。

取得Google Cloud Speech3を既製の音声認識エンジンとして使用して音声を変換します

³<https://cloud.google.com/speech-to-text>

アルゴリズム 1 UGIF エンドツーエンドの説明

```

steps ← retrieve_howto(user_query) マクロ ← parse(steps)

私 ← 0
while i < len(マクロ) do マクロ ← マク
  ロ[i] アクション ← グラウンド(マ
  クロ、スクリーン)
  アクション = SCROLLの場合
    i ← i + 1 終了
  if
    終了する

```

説明文電話アプリを開	マクロシーケンス
きます。[最近] をタップし ます。	tap("Phone"); tap("最 近"); タップ("設定"); タップ
設定アプリを開きます。[ネット ワークとインターネット] をタッ プします。Wi-Fiをオフにします。	("ネットワーク & インターネット"); toggle("Wi-Fi", 間違い);

表 2: サンプル命令と対応するマクロ シーケンス (セクション 4)。

テキストに、多言語埋め込みモデル(Feng et al., 2020)を使用してクエリに対応するベクトルを取得し、それを使用してUGIF-DataSetコーパスのコサイン類似度によって最も類似したハウツーを取得します。

解析解析モデルは、ハウツー命令を受け取り、一連のマクロを生成します (表2)。

PaLM (Chowdhery et al., 2022)、GPT-3 (Brown et al., 2020)、T5 (Raffel et al., 2020)、UL2 (Tay et al., 2022)など、さまざまな言語モデルを試しました。命令を与えられたマクロを生成する文章。

グラウンディング グラウンディング モデルは、現在の UI 画面とともに入力としてマクロ (場合によっては引数を含む) を受け取り、UIで一連のアクションを実行して、マクロによって指定されたタスクを完了します。セットアップのマクロを表1に示します。

tap()とtoggle()の両方について、これらのマクロの引数で参照されている UI 要素を見つける必要があります。つまり、UI 要素と現在画面に表示されている UI 要素のリストを参照する引数を持つマクロが与えられ、どの要素を選択するかを決定する必要があります (または、まったく選択せずにスクロールしてより一致するようにします)。

最も一致する UI 要素を見つけるために、

ジャカードの類似性、UiBERT (Bai et al., 2021)、および多言語 BERT センテンスembedding (LaBSE) (Feng et al., 2020)の実験。UI 要素と参照表現の間のジャカード類似度は、UI 文字列と参照表現の単語を分割し、これら 2 つのセット間のジャカード類似度を見つけることによって測定されます。LaBSE モデルは、センテンス全体の埋め込みを生成します。これを使用して、各 UI 要素の埋め込みと、マクロ内の入力参照式の埋め込みを計算します。参照式と UI要素の埋め込みの内積は、simi のスカラー尺度として使用されます。

マクロに対する引数とUI 要素の間の多様性。スクロールしきい値Tを使用して、現在画面上にある UI 要素をスクロールするか受け入れるかを決定します。類似性メトリックがT未満の場合は、スクロール ダウンしてより一致するものを探します。類似性メトリックがTを超える場合は、最も一致する UI 要素がインタラクション (タップまたはトグル) のために選択されます。Tの適切な値は、開発セットでの実験を通じて決定されます。同様に、UiBERT を使用して、入力参照式とともに画面上のすべての UI 要素の埋め込みを生成しますが、UiBERT では追加の「見つかりません」UI 要素を導入します。

モデルは、スクロール アクションが

取られました。

タップ マクロの場合、マクロ内の引数に最も類似した UI 要素を探すだけで十分です。ただし、トグル マクロの場合、LaBSE 埋め込みを使用する場合、最初にtoggle()マクロへの引数によって参照されるUI 要素を見つけてから、ビュー階層内の近くにある Android Switch 要素を探します (図3)。これは、アプリが標準の Android Switch 要素と、テキスト フィールドが Switch 要素の近くにあるモバイル UI の単純な XML レイアウトを使用している限り機能します。とはいえ、そのようなヒューリスティックは脆弱であり、マルチモーダルモデルによって解決できる可能性があります。

5 実験

UGIF-DataSetデータセットには、ハウツー説明テキストごとに手動で注釈が付けられたオラクル パース (マクロ シーケンス) が含まれています。生成された解析とオラクル解析の間の正確な一致を探すことにより、解析精度を測定します。

データセットには、ハウツー全体の手動で注釈が付けられた画面アクション シーケンスも含まれていますが、

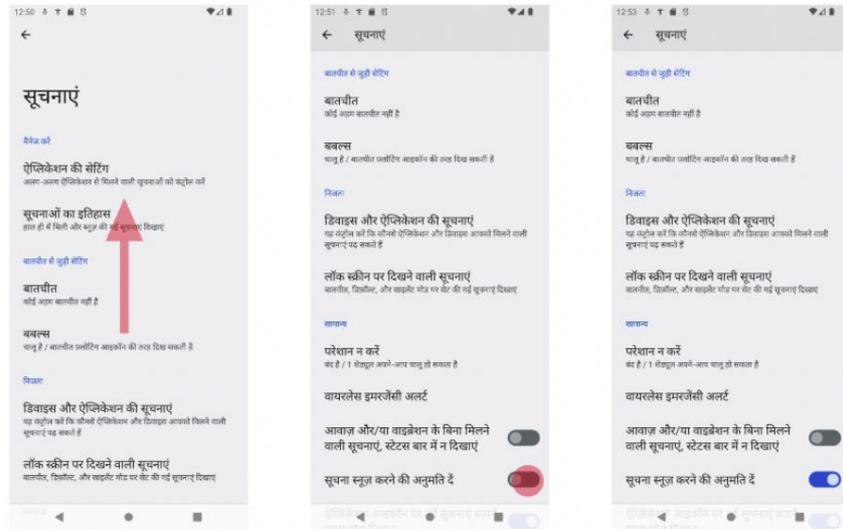


図 3: マクロの実行による UI 画面とアクションのサンプル シーケンス: toggle("Allow notifications snoozing", True)。UI グラウンディング モデルは、どの UI 要素もマクロの引数の文字列に十分に一致していないことを認識し、下にスクロールして一致を見つけ、最も近いスイッチをタップしてオンにします (セクション4)。

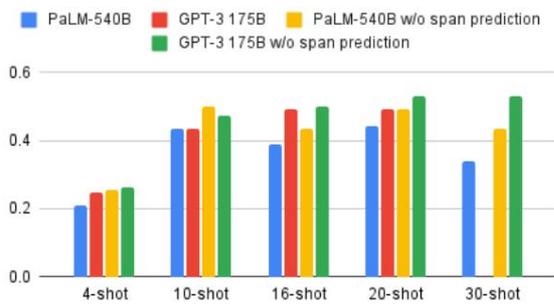


図 4: UGIF-DataSet の開発セットの解析精度 (セクション5.2)。

各マクロにそのようなシーケンスはありません。したがって、グラウンディング モデルを評価するために、エンド ツー エンドのタスク完了率を考慮します。各タスクを複数の方法で完了することは可能ですが、説明テキストのハウツーに正確に従いたいため、モデルによって予測されたアクションのシーケンス全体が正確に一致する場合にのみ、タスクが正常に完了したと見なします。アノテーターが実行する一連のアクション。

5.1 検索は全体でどの程度うまく機能しますか
言語？

多言語文埋め込みモデル(Feng et al., 2020)は、非 EN 言語のハウツー クエリと英語のハウツークエリのマッチングに優れています(表5.1)。EN 以外のテキスト クエリの失敗を調べると、クエリのごく一部が繰り返しであるデータセットのノイズが明らかになりました。

モデル	解析中
構成	精度46%
PaLM 540B 20ショットICL	
GPT-3 175B 20ショットICL	50.9%
PaLM 8B ソフトプロンプトチューン	49.1%
PaLM 62B ソフトプロンプトチューン	64.9%
PaLM 540B ソフトプロンプトチューン	66.8%
UL2 20B フルファインチューン	66.8%
T5 11B フル微調整	66.8%
PaLM 8B フルファインチューン	64.5%
PaLM 62B フルファインチューン	67.5%
PaLM 540B フルファインチューン	70.1%

表 3: UGIF-DataSet テスト セットでの事前トレーニング済みモデル (スパン予測中間ステップなし) の解析精度。インコンテキスト学習 (ICL) は、ランダムに選択された 20 個のトレーニング サンプルを使用します。微調整は、158 個のトレーニング サンプルすべてで実行されます。ソフト プロンプト チューニングは、50 トークンのソフト プロンプト プレフィックスを使用して行われ、すべてのトレーニング サンプルでも実行されます (セクション5.2)。

句読点などのわずかなバリエーション。 Google Cloud Speech API を市販の自動音声認識装置 (ASR) として使用して音声入力をテキストに変換すると、すべての言語でパフォーマンスが大幅に低下しますが、スワヒリ語では大幅に低下します。また、ASRの失敗は音声の明瞭度の低さ、バックグラウンドノイズ、および「キャッシュ」などの専門用語が原因であることがわかりました。

モデル構成	UI 言語	en	kn	mr
	gu	hi	Oracle パース、Jaccard	グラウンド 十億 SW
55.4	-----	Oracle パース、UiBERT	36.6	33.2
41.1	33.7	40.7	49.8	35.4
Oracle パース、LaBSE	グラウ	48.6	33.6	36.6
ンド PaLM 540B	パース、	LaBSE	グラ	48.6
				33.6
				36.6
				38.5
				40
				37.7
				46.4
				32.1

表 4: UGIF-DataSet テスト セット (セクション5.3.1)でのさまざまなモデル構成のエンド ツー エンドのタスク完了成功率。

クエリ	Oracle テキスト	ASR テキスト
言語 P@1	100	P@1
ja		94.4
kn	97.9	88.6
氏	98.1	91.7
ぐ	97.3	89.6
ひ	94.6	91.3
十億	97.3	91.2
SW	93.0	76.4
エス	96.5	94.8

表 5: さまざまな言語のクエリから英語で最も一致するハウツーを取得するパフォーマンスの比較 (セクション5.1)。

5.2 解析のパフォーマンスは、データセットとモデルのサイズによってどのように変化しますか？

4 ショット プロンプトから 10 ショット プロンプトへと解析パフォーマンスが急激に向上しています(図4)。30 例では、入力のトークン数がハードウェアが処理できる最大数を超え、パフォーマンスが低下します。指示テキスト内の顕著な範囲を、一連の思考を促すための中間ステップとしてマークすると(Wei et al., 2022)、解析のパフォーマンスが低下します。利用可能なすべてのトレーニング サンプルを完全な微調整またはソフト プロンプト チューニングで使用すると(Lester et al., 2021)、結果として得られるパフォーマンスは、少数ショットのプロンプトよりも大幅に優れています(表3)。完全な微調整を使用した場合、解析の精度はモデルのサイズによってわずかにしか向上しません。ただし、ソフト プロンプト チューニングを使用すると、より大きなモデルを使用する利点が大きくなります。

5.3 解析のための大規模な言語モデルの一般的な失敗モードは何ですか？

モデルの予測が正しくないテスト サンプルを調べたところ(図5)、微調整された PaLM 540B モデルが (a) 間違っ

1 Text: open the google play app google play. at the top right, tap the profile icon. tap settings and then authentication and then require authentication for purchases.
Code: tap("google play"); tap("settings"); tap("authentication"); tap("require authentication for purchases");

2 Text: on your android phone or tablet, open the gmail app. select one or more emails. in the top right, tap more and then report spam.
Code: tap("gmail"); select("one or more emails"); tap("more"); tap("report spam");

3 Text: On your Android phone or tablet, open the Google Earth app. On the left, tap Map Style. Turn Gridlines on.
Code: tap("Earth"); tap("Map Style"); tap("Gridlines"); toggle("Gridlines", True);

図 5: 20 ショット プロンプトの PaLM 540B モデルによって生成されたマクロの不適切なシーケンス。最初の例では、出力でマクロ tap("profile icon") が省略されています。2 番目の例では、モデルは存在しない select() マクロを幻覚させます。最後の例では、不要なタップを生成しています: tap("Gridlines") (セクション5.3)。

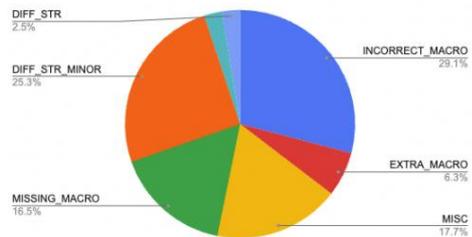


図 6: PaLM 540B 微調整モデル (セクション5.3) によって行われた解析エラーの種類。

(c) 入力命令の重要な部分を見逃してマクロをスキップする、(d) 存在しないマクロを幻覚させる (図6)。

5.3.1 既存のモデルは UI グラウンディングにどの程度有効ですか？

単純な文字列マッチング モデルでも、ハウツーの言語が UI 言語と一致する場合に優れたパフォーマンスを提供できることがわかりました(表4)。驚いたことに、UiBERT はこの基準を下回りました。説明テキストと UI 言語の場合

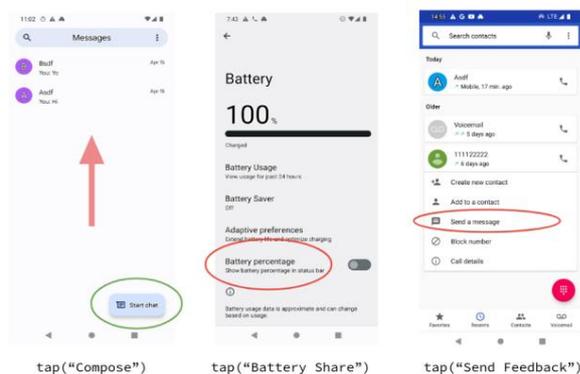


図 7: UI グラウンディング モデルは、UI の状態とマクロが与えられた場合に不適切なアクションを選択します。最初の例では、モデルは「作成」の一致する要素として「チャットの開始」をタップする必要がありますが、代わりに下にスクロールしようとし、一致する UI 要素が見つからないというエラーをスローします。2 番目の例では、モデルは下にスクロールして「Battery share」を見つける必要がありますが、部分的に一致する「Battery percentage」を誤って選択しています。最後の例では、モデルは「フィードバックを送信」ボタンが UI にないことを認識し、エラーをスローする必要がありました。代わりに部分的に一致する「メッセージを送信」ボタンを誤って選択しました (セクション5)。

多言語モデルである LaBSE を使用する必要がありますが、英語でのパフォーマンスは他の言語よりも優れていることがわかりました。LaBSEを使用して誤って予測されたサンプル(図7)を調べると、これらの失敗モードが明らかになりました(図8)。84.5%、(b) モデルはオーバートリガーし、スクロールしてより適切な一致を探す代わりに、不正確な一致を選択します (5.2%)、(c) モデルは一般的な UI パターンとアプリ名の知識を欠いているため、「Google Play」に最も近いものを見つけようとすると、「Play Store」と「Google One」 (5.2%)。

グラウンディング モデルが過度にトリガーされ、部分的に一致する UI 要素が選択され、スクロールダウンに失敗したり、ハウツーが古くなっていることを認識できなくなったりすると、UI で手順が誤って実行されます。これらは、ユーザー エクスペリエンスの低下につながるため、最も深刻な問題です。

さらに、UGIF -DataSetのサンプルの 29% がAndroid 12の UI と一致しない指示テキストを持っているとしてアノテーターによってマークされているという事実によって証明されるように、ヘルプ記事は頻繁に古くなります。

また、PixelHelpデータセットで最高のパフォーマンスを発揮するモデルを評価しました(Li et al., 2020a)。表6

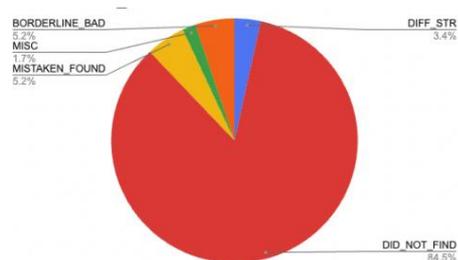


図 8: LaBSE を使用した UI グラウンディング エラーのカテゴリ (セクション5.3.1)。

モデル、データセット	成功 レート
李等。(2020a)、PixelHelp (en) 70.5% 当社、PixelHelp (en) 71.1% 当社、UGIF-DataSet (en)	48.6%
私たちの、UGIF-DataSet (sw)	32.1%

表 6: さまざまなデータセットでの最高のパフォーマンスを発揮するモデル (解析用の PaLM 540B と接地用の LaBSE) の比較。PixelHelp (en) データセットと UGIF-DataSet (sw) のモデルパフォーマンスの間には大きなギャップがあり、改善のためのヘッドルームを検討することを示唆しています (セクション5)。

は、UGIF-DataSetが、特に非 EN 言語で改善の余地が大幅に大きい、より難しいデータセットであることを示しています。

6 結論

新しいスマートフォン ユーザーは、自分の電話の機能を探索するのに苦労しています。音声クエリに基づいて UI でタスクを実行する方法を示すことで、電話をより快適に使用できるようになることをお勧めします。UI言語が指示テキストで使用される言語と異なる可能性があるUIでハウツー指示を取得して実行するタスクについて、既存の言語および文の類似性モデルを評価しました。このタスクのために構築するモデルは、アプリの新しいバージョンが頻繁にリリースされ、構造が古くなるため、UI の小さな変化に適応する必要があります。多言語 UI では、アプリ開発者がすべての UI 要素の翻訳を提供していない可能性があるため、単一の UI 画面で複数の言語を同時に操作しなければならないという課題が生じます。これは、言語固有のモデルではなく、多言語モデルのケースを強化します。最後に、現在の事前トレーニング済みモデルの評価では、大幅な改善の余地があり、マルチモーダル言語 UI 基盤モデルが大幅な利益につながる可能性があることが示唆されています。

参考文献

- アンドロイド。2022. ui automa で自動テストを書く
トル。Android ドキュメント。
- Gilles BaechlerとSrinivas Sunkara。2021. アイコン検出によるモバイルアプリのアクセシビリティの改善。
Google AI ブログ。
- Chongyang Bai,Xiaoxue Zang,Ying Xu,Srinivas Sunkara,Abhinav Rastogi,Jindong Chen,他
2021. Uibert: UIを理解するための一般的なマルチモーダル表現の学習。arXiv プレプリント arXiv:2107.13731.
- Satchuthananthavale RK Branavan,Harr Chen,Luke Zettlemoyer,および Regina Barzilay。2009. 指示を行動にマッピングするための強化学習。
ACL の第 47 回年次総会および AFNLP の自然言語処理に関する第 4 回国際合同会議の議事録,82 ~ 90 ページ。
- SRK Branavan,Luke Zettlemoyer,および Regina Barzi が寝ていました。2010. 行間を読む: 高レベルの命令をコマンドにマッピングすることを学ぶ。計算言語学協会の第 48 回年次総会の議事録,1268 ~ 1277 ページ。
- トム・ブラウン、ベンジャミン・マン、ニック・ライダー、メラニー・サブピア、ジャレッド・D・カプラン、ブラフ・ダリワル、アービンド・ニラカタン、プラナフ・シャム、ギリッシュ・サストリー、アマンダ・アスケル 他
2020. 言語モデルは数ショット学習器です。神経情報処理システムの進歩,33:1877-1901。
- Aakanksha Chowdhery,Sharan Narang,Jacob Devlin,Maarten Bosma,Gaurav Mishra,Adam Roberts,Paul Barham,Hyung Won Chung,Charles Sutton,Sebastian Gehrmann 他2022. Palm: パスウェイによる言語モデリングのスケールアップ。arXiv プレプリント arXiv:2204.02311.
- Fangxiaoyu Feng,Yinfei Yang,Daniel Cer,Naveen Arivazhagan,Wei Wang。2020. 言語にとらわれないパート文の埋め込み。
arXiv プレプリント arXiv:2007.01852.
- Zecheng He,Srinivas Sunkara,Xiaoxue Zang,Ying Xu,Lijuan Liu,Neven Wickers,Gabriel Schubiner,Ruby Lee,Jindong Chen。2021. Actionbert: ユーザー インターフェイスのセマンティックな理解のためのユーザー アクションの活用。人工知能に関する AAAI 会議の議事録,第 35 巻,5931 ~ 5938 ページ。
- ブライアン・レスター、ラミ・アル＝ルファー、ノア・コンスタント。2021年。
パラメータ効率の高い迅速なチューニングのためのスケールの力。
arXiv プレプリント arXiv:2104.08691.
- 魏李。2021. マクロ アクションで構成されたデモンストレーションを通じて UI ナビゲーションを学習します。arXiv プレプリント arXiv:2110.08653.
- Yang Li,Jiacong He,Xin Zhou,Yuan Zhang,Ja son Baldrige。2020a.自然言語の指示をモバイル UI アクション シーケンスにマッピングします。arXiv プレプリント arXiv:2005.03776.
- Yang Li,Gang Li,Luheng He,Jingjie Zheng,Hong Li,Zhiwei Guan。2020b.ウィジェットのキャプション: モバイル ユーザー インターフェイス要素の自然言語による説明を生成します。arXiv プレプリント arXiv:2010.04295.
- Evan Zheran Liu,Kelvin Guu,Panupong Pasupat,Tian lin Shi,Percy Liang。2018. ワークフローに基づく探索を使用した Web インターフェイスでの強化学習。arXiv プレプリント arXiv:1802.08802。
- コーリー・リンチとピエール・セルマネ。2020. 非構造化データに対する言語条件付き模倣学習。arXiv プレプリント arXiv:2005.07648.
- Ramesh Manuvinakurike,Jacqueline Brixey,Trung Bui,Walter Chang,Doo Soon Kim,Ron Artstein,および Kallirroi Georgila。2018. Edit me: 自然言語による画像編集を理解するためのコーパスとフレームワーク。言語資源と評価に関する第 11 回国際会議の議事録 (LREC 2018)。
- コリン・ラフェル、ノーム・シェイザー、アダム・ロバーツ、キャサリン・リー、シヤラン・ナラン、マイケル・マテナ、ヤンチー・チョウ、ウェイ・リー、ピーター・J・リユ 他2020. 統合されたテキストからテキストへのトランスフォーマーを使用して転移学習の限界を探る。J. マッハ。学び。決議,21(140):1-67。
- ピーコシュランジャン。2022.よりインクルーシブなインターネットのための洞察のアンソロジー。Google ブログ。
- ジャンナ・サーセンバエフ。2018. モバイルインタラクション中の状況障害。2018 ACM International Joint Conference および 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers の議事録,498 ~ 503 ページ。
- Tianlin Shi,Andrej Karpathy,Linxi Fan,Jonathan Hernandez,および Percy Liang。2017. World of Bits: Web ベースのエージェント向けのオープン ドメイン プラットフォーム。機械学習に関する国際会議,3135 ~ 3144 ページ。PMLR。
- Yi Tay,Mostafa Dehghani,Vinh Q Tran,Xavier Garcia,Dara Bahri,Tal Schuster,Huaixiu Steven Zheng,Neil Houlsby,Donald Metzler。2022. 言語学習パラダイムの統一。arXiv プレプリント arXiv:2205.05131.
- Daniel Toyama,Philippe Hamel,Anita Gergely,Gheorghe Comanici,Amelia Glaese,Zafarali Ahmed,Tyler Jackson,Shibl Mourad,Doina Precup。2021. Androidenv: Android 向けの強化学習プラットフォーム。arXiv プレプリント arXiv:2105.13231.
- Sagar Gubbi Venkatesh,Anirban Biswas,Raviteja Upadrashta,Vikram Srinivasan,Partha Talukdar,Bharadwaj Amrutur。2021. 空間推論

ロボット操作のための自然言語命令から。2021年のIEEE International Conference on Robotics and Automation (ICRA)、ページ 11196 ~ 11202。IEEE。

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, Denny Zhou。2022年。
思考を促すチェーンは、大規模な言語モデルで推論を引き出します。arXiv プレプリント arXiv:2201.11903。

Nancy Xu, Sam Masling, Michael Du, Giovanni Campagna, Larry Heck, James Landay, Monica S Lam。2021。ウェブ サポート タスクを自動化するためのオープン ドメインの手順の接地。arXiv プレプリント arXiv:2103.16057。

Mingyuan Zhong, Gang Li, Peggy Chi, Yang Li。
2021。Helpviz: テキストベースの指示からコンテキストに応じた視覚的なモバイル チュートリアルを自動生成。
ユーザー インターフェイス ソフトウェアとテクノロジーに関する第 34 回年次 ACM シンポジウム、1144 ~ 1153 ページ。