論文要約

LLM関連

Instruction Embedding: Latent Representations of Instructions Towards Task Identification 指示埋め込み：タスク識別に向けた指示の潜在表現

概要: 指示データを分全体の意味ではなくタスク特定に重点を置くために動詞　-名詞ペアを強調する Instruction Embeddingを提案しタスク識別を行い

この埋め込みを LLMのプロンプトを使用し特定のタスクに注意を向けるようにする、　PIE手法を利用する。

従来の研究では、指示のテキスト埋め込みは主に全体的な意味情報の取得に焦点が置かれていましたが、この論文では、指示埋め込みはタスクの特定に重点を置くべきであると指摘しています。従来の埋め込み手法では、異なるタスク間の意味的類似性が強調され、異なるタスクを正確に区別することが難しかったが、本研究の　PIE手法はこの問題に対処し、より精度の高いタスク特定を可能にしています。

技術や手法の説明：

1. **指示埋め込み（Instruction Embedding）**:
   - 文全体の意味的情報ではなく、タスク特定に重点を置いた埋め込み方法。タスクの動詞　-名詞ペアを強調し、タスクの同定をより正確に行うためのもの。
   - 指示のタスクを識別するため、意味的な類似性よりもタスクの類似性を優先。

2. **IEBベンチマーク（Instruction Embedding Benchmark）**:
   - タスクの識別能力を評価するために構築されたベンチマークデータセット。従来のテキスト埋め込みベンチマークが意味的類似性を評価するのに対し、　IEBはタスクの違いに基づいて指示を分類することを目指す。
   - 47,161のサンプルを含み、1,353のタスクカテゴリーに分類。

3. **PIE（Prompt-based Instruction Embedding）**:

MemSim: A Bayesian Simulator for Evaluating Memory of LLM-based Personal Assistants MemSim: LLMベースのパーソナルアシスタントのメモリを評価するためのベイジアンシミュレーター

概要: LLMエージェントの MemSimではQAにBayesian Relation Network (BRNet) 使用しユーザーのプロファイルを関係性や年齢などの属性でサンプリングし質問に必要な情報をヒントの形でまとめ Aを生成します

**技術や手法 **

1. **Bayesian Relation Network (BRNet):**

ユーザーの属性や関連エンティティをベイジアンネットワークとして表現し、階層的なユーザープロファイルをサンプリングする手法。 BRNetは、ユーザーの関連する属性やエンティティの確率分布をモデル化し、シナリオに応じた多様でスケーラブルなユーザープロファイルを生成する。

2. **Causal Generation Mechanism:**

階層的なユーザープロファイルに基づいて、ユーザーメッセージと QAを因果的に生成するメカニズム。これは、プロファイルからヒント(factual hints)を抽出し、それに基づいてメッセージや QAを作成するプロセスで、LLMのハルシネーションの影響を軽減する。

3. **MemDaily Dataset:**

MemSimを用いて日常生活シナリオを元に構築されたデータセット。シンプルな QAから多段階の推論を必要とする QAまで、さまざまなタイプの QAを含む6つのサブデータセットを提供する。

**使用用途 **

SELF-CONTROLLER: CONTROLLING LLMS WITH MULTIROUND STEP-BY-STEP SELF-AWARENESS セルフコントローラー：マルチラウンドのステップバイステップの自己認識による　LLMの制御

概要：自己認識を導入して LLMが自身の状態と現状を認識し制御性を強化。テキスト長の線形性と単調性に基づいた二分探索アルゴリズムを実装しています

**技術や手法 **

1. **Self-controllerの全体像：**

　Self-controllerは、LLMが自身の出力に基づいて状態を認識し、次の出力を調整するマルチラウンドの対話フレームワークです。これにより、　LLMは自己認識を持ち、より精密な制御が可能になります。

2. **状態反映器（ State Reflector）の役割：**

　- **状態更新：** LLMの各応答を解析し、現在の状態変数を更新します。例えば、テキスト生成タスクでは、これまでに生成した単語数を計算します。

　- **フィードバック提供 ：** 更新された状態を LLMにフィードバックし、次の出力に反映させます。

3. **マルチラウンド対話の流れ ：**

　- **初期プロンプト：** ユーザーからのタスク指示と目標状態（例：目標とするテキスト長）を　LLMに提示します。

　- **応答生成：** LLMは現在の状態と目標に基づいて出力を生成します。

　- **状態更新とフィードバック ：** 状態反映器が LLMの出力を解析し、状態を更新してフィードバックします。

REGENESIS: LLMS CAN GROW INTO REASONING GENERALISTS VIA SELF-IMPROVEMENT REGENESIS: LLMは自己改善を通じて推論の汎用性を持つことが可能になる

概要: ReGenesisは特定タスクに適応する形で推論ガイドラインを生成し、それを使用して推論構造を生成し解決ステップを決定。ステップに基づき回答を生成、生成された推論経路の中から正解の経路をフィルタリングし、これをモデルの再学習に使用します

ReGenesisは既存の自己生成手法（例： STaRなど）が抱える問題を克服しており、特にタスク間の汎用性（ OODタスク）において優れたパフォーマンスを発揮します。従来手法が特定タスクに偏った推論経路 を生成することにより外部タスクでの性能が低下するのに対し、 ReGenesisは一般的なガイドラインを用いることで幅広いタスクに対応可能な推論経路を生成します。

技術や手法

1. **推論ガイドラインの適応 (Guidance Adaption)**: 一般的な問題解決戦略を特定タスクに適応する形で推論ガイドラインを生成する。この段階では問題を直接解決するのではなく、全体的な解決戦略をタスクに適応させる。

2. **推論構造の生成 (Reasoning Structure Generation)**: 適応された推論ガイドラインに基づいて詳細な推論構造を生成します。この段階でも具体的な解を出すことはせず、解決のためのステップを決定し ます。

3. **推論経路の生成 (Reasoning Paths Generation)**: 推論構造に従い、実際の解決手順を生成します。この手順により具体的な答えを得ます。

4. **推論経路のフィルタリング (Filtering Reasoning Paths)**: 生成された推論経路の中から正解に一致する経路をフィルタリングし、モデルの再学習に使用します。

使用用途

ReGenesisは、数学的推論、論理的推論、常識的推論などの複雑な推論タスクに適用可能です。特に、汎用的な推論能力を向上させることにより、未知のタスク（ OODタスク）にも対応可能な LLMを作り出すことができます。

Comparing Criteria Development Across Domain Experts, Lay Users, and Models in Large Language Model Evaluation ドメインエキスパート、一般ユーザー、およびモデル間の大規模言語モデル評価における基準開発の比較

概要: LLMをドメイン固有のタスクで使用した際の評価プロセスを提案

基準設定として事前段階 (a priori)を参加者がプロンプトのみで基準設定、実際の出力を確認し修正する事後段階 (a posteriori) の2段階で設定

これにより専門性とユーザーの理解しやすさ、評価基準の修正ができます

技術や手法

1. **基準設定プロセスの比較 **:

    1. **基準設定の 2段階プロセス **:

        - **事前段階 (a priori)** と **事後段階 (a posteriori)** の2段階に分けて評価基準を設定するプロセスを行います。

        - **事前段階 (a priori)** では、参加者(ドメインエキスパートや一般ユーザー)がプロンプトのみを見て基準を設定します。この段階は、モデルの出力を見ないでどのような基準が必要かを考えるフェーズです。

        - **事後段階 (a posteriori)** では、参加者が実際のモデルの出力を確認した後、基準を修正または追加するフェーズです。この段階は、モデルの実際の出力の内容に基づいて、最初の基準をどのように調整するかを考えるフェーズです。

    2. **参加者グループの分類と役割 **:

        - 研究では、3つの異なるグループが評価に参加しています。

            - **ドメインエキスパート **(栄養学や教育学の専門家): 彼らは深い知識を持ち、具体的な基準を設定することで、特定のドメインでの信頼性を確保する役割を担っています。

            - **一般ユーザー **: 専門的な知識がなく、ユーザー視点から出力の使いやすさや理解しやすさを重視して基準を設定します。

            - **LLM**(大規模言語モデル): モデル自身が基準を生成し、人間の基準と比較する対象となります。

DreamGarden: A Designer Assistant for Growing Games from a Single Prompt DreamGarden: シングルプロンプトからゲームを成長させるデザイナーアシスタント

概要: DreamGardenは、ゲーム環境を作成するデザイナー支援で、プロンプトを階層的に分割し、具体的なアクションプランを生成。プランはサブモジュールに分配し具体的な実装をします。特に Unreal Engineでのゲーム開発を支援し、自由形式のプロンプトから自律的にゲームのプロトタイプ生成します

**技術や手法 **

DreamGardenの中核は、大規模言語モデル(LLM)を使用したプランニングモジュールである。このプランニングモジュールは、ユーザが提供したプロンプトを元に階層的なアクションプランを生成し、それを様々なサブモジュールに割り当てることで実装を行う。具体的には、以下の手法を用いる :

1. **プランニングモジュール **:
   - ユーザが提供するシードプロンプトを元に、高レベルのゲームデザインを広範囲に計画する。
   - その後、各プランを詳細なステップに分割し、最終的に具体的なタスクにまで落とし込む。
2. **実装サブモジュール **:
   - サブモジュールには、 C++コード生成、手続き型メッシュ生成、ディフュージョンメッシュ生成などがある。
   - 各タスクは特定のサブモジュールに割り当てられ、それに応じた実装が行われる。
   - 生成されたコードやアセットは Unreal Engineに取り込まれ、シミュレーションやフィードバックが行われる。
3. **ユーザーインターフェース **:

AI AS HUMANITY'S SALIERI: QUANTIFYING LINGUISTIC CREATIVITY OF LANGUAGE MODELS VIA SYSTEMATIC ATTRIBUTION OF MACHINE TEXT AGAINST WEB TEXT 人類のサリエリとしての AI：ウェブテキストに対する機械テキストの体系的な帰属を通じた言語モデルの創造性の定量化

概要: LLMの人間の創造性評価として CREATIVITY INDEXを提案。ウェブテキストをどの程度再構成できるかを計測

テキストの n-gramがウェブ上に存在するかどうかを評価し厳密一致や近似一致をみる L-ユニークネスを算出します

技術や手法の詳細説明

この論文では、LLM（大型言語モデル）の創造性を測定・評価するために、新しい指標「 CREATIVITY INDEX」と動的プログラミングアルゴリズム「 DJ SEARCH」を提案しています。以下、技術と手法の詳細を順を追って説明します。

1. **CREATIVITY INDEX**

CREATIVITY INDEXは、LLMが生成したテキストの創造性を定量的に測定するために提案された指標です。主なアイデアは、生成されたテキストがどの程度ウェブ上の既存テキストから再構築できるかを評価することです。具体的には、与えられたテキスト内の n-グラム（連続する n個の単語）が、ウェブ上の巨大な参照コーパスに存在するかどうかを確認します。

- **L-ユニークネス (L-uniqueness)**: テキストの中で、指定された長さの n-グラムが参照コーパスに含まれていない割合を示します。これは、テキストの中のどの単語やフレーズが新しい文脈で使われているかを定量的に示すもので、より高い値がより高い創造性を意味します。

- **CREATIVITY INDEXの計算**: テキストのすべての n-グラムについて L-ユニークネスを計算し、それを積み上げて総合的な創造性を示す指数とします。これは、言い換えると「 L-ユニークネス曲線の下の面積」を求めることと同等です。

この指標を用いることで、生成されたテキストがウェブ上の既存のテキストからどの程度新しいものを取り入れているかを定量的に評価し、 LLMの創造性を人間の作家と比較することができます。実験の結果、プロの作家の創造性が LLMよりも大幅に高いことが確認されました。

Document-level Causal Relation Extraction with Knowledge-guided Binary Question Answering 知識ガイド付き二項質問応答による文書レベルの因果関係抽出

概要: Knowledge-guided Binary Question Answering（KnowQA）は因果関係を分類するイベント間因果関係抽出（ ECRE）をLLMのゼロショットで文書内のイベントを抽出しその構造を構築するイベント構造構築と因果関係の有無を識別する二択で答えられる質問を使い分類します

技術や手法

1. イベント構造の構築（ Event Structure Construction Module）

イベント間の因果関係を抽出するために、 KnowQAでは文書レベルでのイベント構造の構築を行います。これは 3つの主要なステップで構成されています。

1.1 イベント検出（ Event Detection）

まず、文書中のイベントを検出します。これは、事象（イベント）を特定し、それを特定のイベントタイプに分類するプロセスです。具体的には、以下の手法が使用されています。

- **KAIROSオントロジー **を利用し、イベントの分類を行います。 KAIROSオントロジーは ACE 2005（Automatic Content Extraction）の拡張セットであり、 50のイベントタイプと 59の引数役割をカバーしています。

- **CLEVE**（Contrastive Pre-training for Event Extraction）と呼ばれる事前学習モデルを使用し、イベントの分類精度を向上させています。このモデルは事前学習された PLM（Pre-trained Language Model）で、**WikiEventsデータセット **でトレーニングされており、各イベントを KAIROSオントロジーに従って分類します。

1.2 イベント引数抽出（ Event Argument Extraction）

Give me a hint: Can LLMs take a hint to solve math problems? ヒントを教えて： LLMは数学の問題を解くためのヒントを活用できるか？

概要: LLMに対してヒントを与えることでの数学問題解決能力についての有効性を評価

モデルに対して質問をヒントや例なしで回答した結果ベースとし、品とありワンショットで例をフューショットで複数例、誤ったヒントやランダムなヒントに対して評価を実施し、ヒントありのものが CoTより良い結果、ワンショットは複雑な問題の解法が難しく、フューショットは与えた例に結果が依存する。誤ったヒントや例を渡すと結果は大幅に悪くなる結果になり、ヒントが数学的推論向上に有効なことがわかりました

評価結果の詳細

この論文では、様々な種類のプロンプト手法（ベースライン、ヒント、ワンショット、フューショット、チェイン・オブ・ソートなど）を使って、 LLMが数学問題を解く能力を評価しています。ここでは、各手法の評価結果について順番に詳しく説明します。

1. ベースラインプロンプト

**ベースライン **では、モデルに対して単に問題を与え、ヒントや例を提示せずに問題を解くよう促しました。結果として、ベースラインのスコアは比較的低く、平均的なパフォーマンスとなりました。これは、ヒントや例がない状態ではモデルが正確に問題を解決するのが難しいことを示しています。

2. ヒントプロンプト

**ヒントプロンプト **を使用した評価では、各問題に対して適切なヒントを与えることで、モデルのパフォーマンスが向上しました。この手法は、人間が数学の問題を解く際に適切な助言を受けることに似ており、モデルにとって有益でした。具体的には、モデルが正しい解法にたどり着くための方向性を持ち、計算ミスや論理的な誤りを減少させる効果が観察されました。この結果、他の手法（特にチェイン・オブ・ソート）よりも優れたスコアを示しました。

MoDEM: Mixture of Domain Expert Models MoDEM: ドメインエキスパートモデルの混合

概要: ドメインプロンプトとドメイン特化したモデルを組み合わせることで汎用モデルに比べ LLMの性能効率が向上する MoDEMを提案。DeBERTa-v3-largeで入力内容をドメイン分類し、健康、数学、科学、コーディングなど各ドメインに特化したモデルを使用して回答します

技術や手法

1. **BERTベースのルーター **

- **使用するモデル **: MoDEMのルーティングシステムでは、 **Microsoft DeBERTa-v3-large**モデルが用いられています。このモデルは、 304Mパラメータを持つ BERT系のモデルで、文分類タスクに特化しています。

- **役割**: このルーターは、入力されたプロンプトを適切なドメインに分類する役割を果たします。ドメインには、数学、健康、科学、コーディング、その他のカテゴリが含まれています。ルーターはこれらのドメインのいずれかを識別し、プロンプトを特定のドメインモデルに送信します。

- **ファインチューニング **: ドメインを分類するために、このルーターは事前に選定されたデータセットを用いてファインチューニングされています。ファインチューニング時には、 1エポックでバッチサイズ 32、学習率1e-5という設定が使用されました。

- **特徴**: ルーターは非常に軽量で、最大の専門モデルの 0.42%のサイズしかないため、リソースの消費が非常に少ないです。分類の精度はテストデータで 97%を達成し、MMLUのようなアウトオブディストリビューションのデータにも高い精度で対応できることが確認されています。

2. **ドメイン専門モデル(Expert Models)**

- **専門モデルの選定 **: MoDEMのドメイン専門モデルは、それぞれのドメインで高性能を発揮するオープンソースモデルを選定しています。これにより、特定のドメインに最適化されたモデルを使って、汎用モデルよりも優れたパフォーマンスを実現しています。

- **中規模モデルセット **:

DA-Code: Agent Data Science Code Generation Benchmark for Large Language Models DA-Code: 大規模言語モデルのためのエージェントデータサイエンスコード生成ベンチマーク

概要: LLMをエージェントベースのデータサイエンスタスクをコード生成の観点で評価するためのベンチマーク、 DA-Codeを提案

データワークリング (DW)、機械学習 (ML)、探索的データ分析 (EDA) の3つのカテゴリについてどれだけ自律的な問題解決ができるかを評価します

1. タスクの構成と分類

DA-Codeは、LLMがデータサイエンスエージェントとしてどれだけの能力を持つかを評価するために、以下の 3つの主要なカテゴリのタスクから構成されています:

- **データワークリング (DW)**: 生データを解析可能な形にするために変換・統合・クリーニングを行う。具体的には、データの読み込み、欠損値の処理、データのクリーニングや統合などが含まれる。

- **探索的データ分析 (EDA)**: データセットの特性を理解し、洞察を得るためのデータ分析を行う。 SQLやPythonを使って統計分析、データマニピュレーション、データの視覚化を行う。

- **機械学習 (ML)**: 機械学習モデルを使って、データに基づく予測や分類を行う。データの前処理から特徴量エンジニアリング、モデルのトレーニング、予測までを実施する。

これらのタスクは全て実際のデータに基づき、 500の複雑なタスクを提供しており、データサイエンスの全てのプロセスをカバーしています。

2. タスクの設計と難易度

- **リアルなデータシナリオ **: DA-Codeのタスクは実際のデータセットを基にしており、単なるノートブック環境でのデータ分析にとどまらず、複数のファイルやデータソースを使用します。例えば、データベースやスプレッドシート、文書、コードなど、複数の情報源からなる多様なデータを使って課題を解決します。

- **タスク難易度の分類 **: タスクは容易、中程度、困難の 3つの難易度に分類されており、それぞれの難易度で異なる種類の技術的な課題が含まれています。具体的には、困難なタスクにはより高度なプロ

AI-Press: A Multi-Agent News Generating and Feedback Simulation System Powered by Large Language Models AI-Press: 大規模言語モデルを活用したマルチエージェントによるニュース生成とフィードバックシミュレーションシステム

概要: 複数のエージェントが協力し合う形でニュースの自動作成と修正を行う　AI-Pressを提案

ニュース生成をドラフティング（草案作成）、ポリッシング（内容修正）、シミュレーション（公開後の反応予測）という　3つのモジュールに分け、各モジュールに複数のエージェントをネット検索や　RAGを使いながらニュース制作を行います

AI-Pressの技術と手法の詳細な説明

AI-Pressは、ニュースの生成からフィードバックのシミュレーションまでを行う自動化システムで、複数のエージェントが協力する形で効率的にニュース制作をサポートします。

1. マルチエージェントシステム

AI-Pressは、ニュース生成のプロセスを「ドラフティング（草案作成）」「ポリッシング（内容修正）」「シミュレーション（公開後の反応予測）」という　3つのモジュールに分け、各モジュールに複数のエージェントを配置して効率化を図っています。このマルチエージェントシステムにより、各エージェントが特定の役割を担い、ニュース制作の精度と効率が向上します。

- **ドラフティングモジュール **: ニュース草案を作成するプロセスで、 Searcherエージェントと Writerエージェントが協力します。

  - **Searcherエージェント **: 提供されたトピックや素材に基づき、多次元的な情報を検索・収集します。このエージェントは、ニュースデータベース、ファクトデータベース、インターネットから情報を収集し、精度と信頼性を確保します。

  - **Writerエージェント **: Searcherエージェントが収集した情報をもとに、ニュース記事の草案を作成します。異なるニュースジャンル（ニュース、プロフィール、コメンタリー）ごとに特化した書き方を持ち、プロフェッショナルな内容に仕上げることを目指します。

- **ポリッシングモジュール **: 初期の草案をさらに洗練させるプロセスで、 Reviewerエージェントと Rewriterエージェントが関与します。

Appendix