

論文要約

LLM関連

概要: LLMを使用して物語が目標語彙の意味をどれだけよく伝えているかを自動評価

180の子ども向け物語を使い 5つのターゲット語を予測、人間の判断とのスピアマン相関は RoBERTaが0.4601でGeminiが0.4983を達成

技術や手法の詳細

1. **データセット**:

- 自動生成された 180の子ども向け物語が対象。
- 物語には5つのターゲット語が含まれ、これらの文脈的サポートが評価された。

2. **文脈的情報量の評価基準**:

- ターゲット語が提供する文脈的情報を、コサイン類似度を用いてスコア化。
- ConceptNet Numberbatch 19.08の単語埋め込みを利用し、ターゲット語と予測語の類似度を計算。

3. **モデル**:

- **RoBERTa**: ターゲット語の予測と埋め込み類似度計算に基づくアプローチ。
- **Gemini**: GoogleのLLMを活用した最先端モデル。特定のプロンプトを使用してターゲット語の予測を行う。

4. **評価結果**:

- Geminiモデルはスピアマン相関 0.4983を達成し、RoBERTa(0.4601) やその他ベースラインを上回る。

5. **一般化能力**:

概要: LLMで生成した複数のコードから正確なコード選択するためのコード識別器である Condorを提案

テキスト的には類似していても機能的に異なるコードの差異を埋め込みレベルのコントラスト学習で識別。バグ修正プロセスで得られる部分的な修正データを使い中間コード拡張を行うこと得 dコードの違いを捉え識別を行います

技術や手法

1. **埋め込みレベルのコントラスト学習**

- テキスト的に類似しているコードペア (正解 / 不正解) を利用。
- ユークリッド距離に基づく損失関数を設計し、正しいコードの埋め込み間の距離を縮小し、不正なコードとの差異を拡大。
- 損失関数:

$$L = \sum_{i=1}^N [l_{a,b} \cdot d(x_a, x_b)^2 + (1 - l_{a,b}) \cdot \max(0, m - d(x_a, x_b))^2]$$

ここで $l_{a,b}$ はラベル、 $d(x_a, x_b)$ はユークリッド距離、 m はマージン。

$$L = \frac{1}{N} \sum_{i=1}^N [l_{a,b} \cdot d(x_a, x_b)^2 + (1 - l_{a,b}) \cdot \max(0, m - d(x_a, x_b))^2]$$

$$L = \frac{1}{N} \sum_{i=1}^N \left[l_{[a,b]} \cdot d(x_{[a]}, x_{[b]})^2 + (1 - l_{[a,b]}) \cdot \max(0, m - d(x_{[a]}, x_{[b]}))^2 \right]$$

$$l_a, l_b \in \{a, b\}$$

$$d(x_a, x_b) = d(x_{[a]}, x_{[b]})$$

概要: 構造化画像理解は、画像内の異なる構造やテキスト間で戦略的に焦点を移し、推論のシーケンスを形成して最終的な答えるために、REFOCUSフレームワークを提案。LLMに画像編集ツールを使用して「視覚的思考」を生成させる能力を付与。画像を編集して視覚的推論プロセスを強化。

1. **選択的注意と多段階推論の欠如を補完 **

既存のLLMはテキスト形式でのみ推論を行うが、REFOCUSは画像編集を通じて選択的注意を視覚的に実現する。

2. **視覚的推論プロセスの改良 **

従来の手法(Visual Sketchpadなど)は自然画像や外部情報に依存しているが、REFOCUSは構造化画像を対象とし、追加の情報を必要とせずに推論能力を向上させる。

提案手法と技術

****REFOCUSフレームワークの概要 :****

概要: FairCodeは、コード生成における LLMの社会的バイアスを評価する新たなベンチマークです。 2つの主要なタスク(関数実装とテストケース生成)を通じて、さまざまなシナリオでモデルのバイアスを評価します。この研究では、モデルのバイアスを定量的に評価する新しい指標「 FairScore」を提案しました。複数の LLMを用いた実験では、全てのモデルがバイアスを示しており、社会的属性(性別、年齢、人種など)に依存する結果が観察されました。

- **先行研究の限界:** 既存の研究は、差別的なプロンプトや、識別モデル向けのデータセットを再利用する方法に依存していました。しかし、生成モデルの進化に伴い、これらの方法では効果的にバイアスを評価できません。

- **新規性:** FairCodeは、生成モデル特有の課題を考慮した専用のベンチマークと指標を提供し、より包括的な評価を可能にしています。また、テストケース生成におけるバイアスの評価は本研究が初めてです。

技術や手法

1. FairCodeの構造と評価タスク

FairCodeは以下の2つの主要タスクから構成されます。

1.1 関数実装

このタスクでは、LLMに与えたプロンプトに基づき関数を生成させ、その中で社会的に敏感な属性(性別、年齢、人種など)が結果に影響を及ぼすかを評価します。

Agent Laboratory: Using LLM Agents as Research Assistants エージェントラボラトリー : LLMエージェントを研究助手として使用する

概要: 研究を文献レビュー、実験計画と実施、レポート作成の 3段階のフィードバックを基に格プロセスで人間の補助ツールとして機能します

1. 文献レビュー (Literature Review)

概要:

この段階では、研究アイデアに基づいて関連する文献を収集・要約し、次の段階で利用するための文献リストを構築します。

技術と手法:

1. **PhDエージェント**:

- 文献検索には arXiv APIを使用。
- 3つのアクションを実行:
 - **Summary**: トップ20件の論文の要約を取得。
 - **Full Text**: 特定の論文の全文を抽出。
 - **Add Paper**: 要約または全文をレビューリストに追加。
- 複数回の検索と評価を繰り返し、関連性の高い文献を選定。

概要: 3つのLLM(GPT-4 Turbo、Claude 3、Mixtral 8x22B)と4つのプロンプト技術を使用し、 12年生向けのテキストを 8年生、6年生、4年生向けに簡略化する方法を評価。学年レベルへの適合性、キーワード と キーフレーズの一貫性、語数の変化率を評価指標を使い、

6年生レベルおよび 8年生レベルで CoTが6年生レベルで DSP、PCで目標レベルの正確性と一貫性を達成しましたが、 4年生レベルでは達成できませんでした

技術や手法

1. **プロンプト技術**:

- **ゼロショット**: 手本なしでのプロンプト実行。
- **Directional Stimulus Prompting (DSP)**: キーワードとフレーズを明示的に維持。
- **Chain-of-Thought (CoT)**: 複雑な推論を中間ステップを経て解決。
- **Prompt Chaining (PC)**: タスクを小分けにした連続プロンプト。

2. **多エージェントアーキテクチャ**:

- **マネージャーエージェント**: キーワード選定やタスク割り振りを実行。
- **ライターエージェント**: テキストの簡略化。
- **エディターエージェント**: 結果のレビューと修正提案。
- **計算エージェント**: 読みやすさや語数の計算。

3. **評価指標**:

技術や手法

1. IRAGモジュール

- **構成意図の抽出**:
 - LLMを用いて構成を機能別に分割し、意図を抽出。
 - テンプレートや例を活用した「 In-Context Learning」により意図抽出の統一性を向上。
- **ターゲットマニュアルの検索**:
 - LLMを用いたマニュアルフィルタリングと投票メカニズムにより検索精度を向上。
 - BERTベースのモデルを使用し、意図とマニュアルの類似性を計算。
- **インクリメンタル翻訳**:
 - コンテキストを保持しながら断片的に翻訳を進める。
 - 翻訳ごとに前回の結果を利用して文脈依存性を考慮。

2. 二段階検証モジュール

- **構文検証**:

概要: 現音声アシスタントで長期記憶を効率的に管理するため、事前定義されたカテゴリに基づくユーザーの好みの抽出、保存、取得を可能にするシステムを提案

プライバシーの懸念や規制への対応を考慮しながら、LLMで好みを抽出して記憶

技術や手法

1. **好みの抽出 (Preference Extraction)**

- **カテゴリ境界に基づく抽出**:

- 事前に定義された階層型カテゴリに基づき、ユーザーの発言から好みを抽出。
- LLMの関数呼び出しを用いて、JSON形式で構造化された情報を生成。
- 外部カテゴリ情報は無視し、ユーザーの選択に応じた抽出も可能。

2. **好みのメンテナンス (Preference Maintenance)**

- 重複や矛盾を削減するため、3つのメンテナンス関数(追加、更新、保持)を実装。
- カテゴリごとに許容される好みの数(単一または複数)を制御。

Appendix
