


Document AI タスクに向けた大規模事前学習済みモデルを活用した Layout-aware Prompting

北田 俊輔¹, 井上直人², 大谷まゆ², 彌富仁¹  法政大学
¹法政大学 理工学研究科 応用情報工学専攻 ²株式会社CyberAgent

Summary

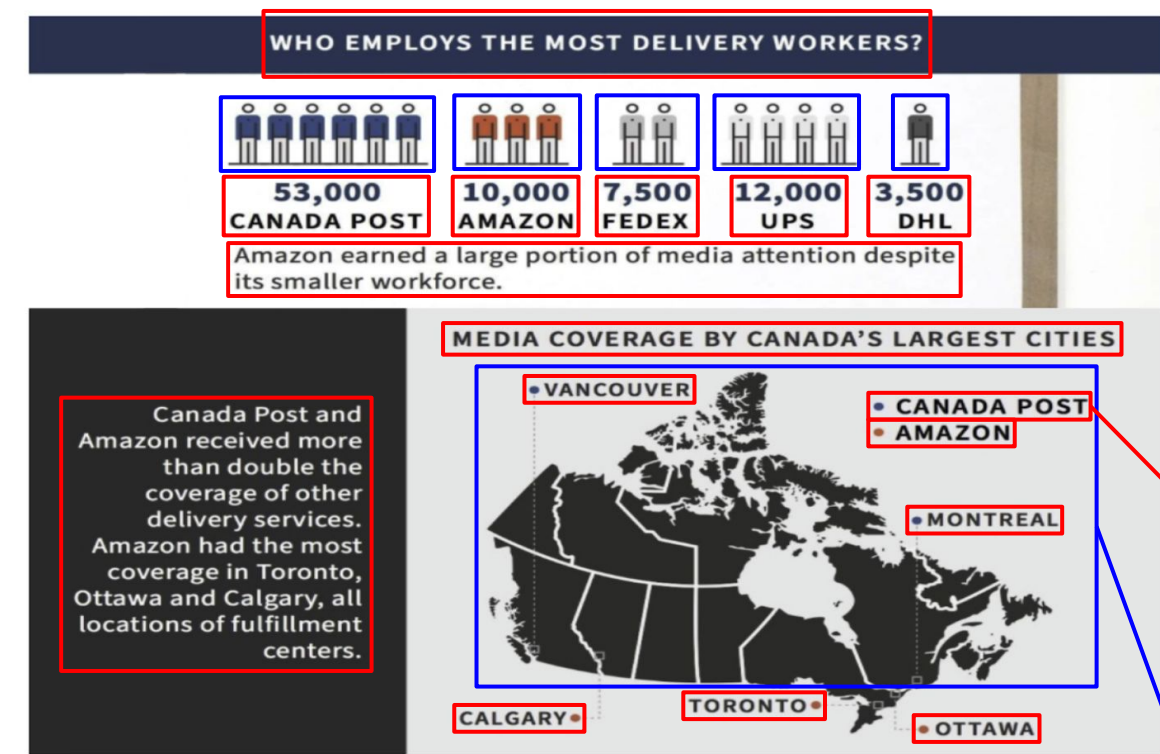
大規模事前学習済み言語モデルを Document AI タスクへ効率的に応用する際のレイアウト情報を考慮した新たな prompting 技術の提案

- ✓ 既存の GPT モデルに対してテキストのレイアウト情報を埋め込んだ layout embedding を加えるだけの非常にシンプルかつ効果の高い手法
- ✓ テキストや画像を始め、イラストやグラフといったオブジェクトを含むインフォグラフィック質問 応答タスクを例に、提案手法の効果を確認

Background

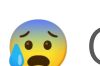
インフォグラフィック理解: 主にビジネス文書に焦点を当てた Document AI タスクの一つ

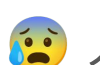
- 一般物体ではなく人工物体を理解する必要あり
 - テキストとイラスト間の位置の関係性の把握が重要
- Webページのスクショ等に対する情報抽出等に応用可能



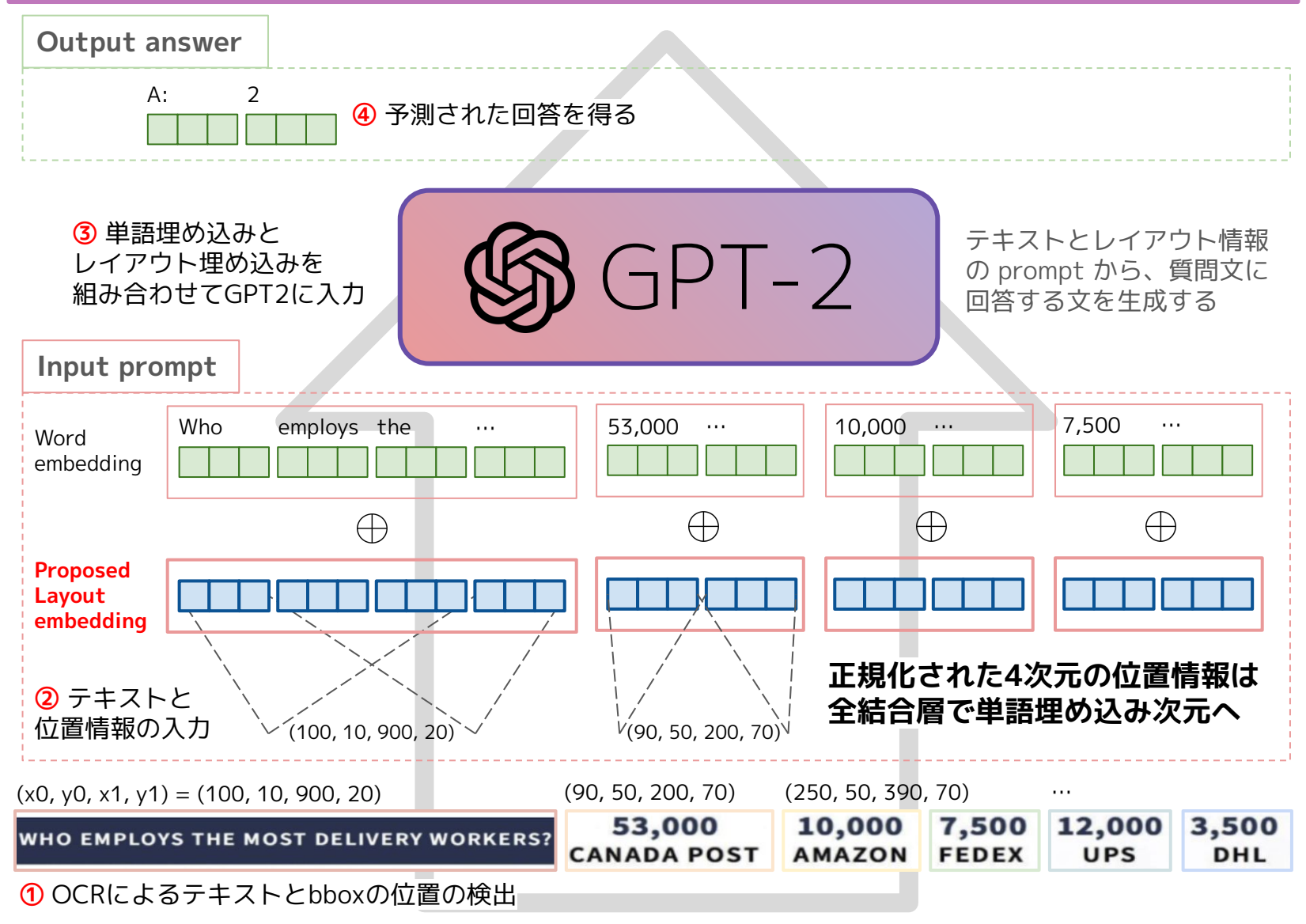
How many companies have more than 10K delivery workers?
 Answer: 2 Evidence: Figure
 Answer-source: Non-extractive Operation: Counting Sorting

インフォグラフィックに対する様々な質問へ回答するタスク

(1) OCR してテキスト取得
 OCR 誤り

(2) イラストから物体認識
 人工物体

Layout-aware Prompting (LAP)



Experiments

評価用データセット

- Infographic VQA [Mathew+ WACV'22] ([コンペwebサイト](#))
 - 30,035 質問応答対 - 5,485 画像
 - 先行研究のVQAデータセットより質問文が長く複雑

比較手法

- ベースライン、提案手法、SoTa モデル

- **GPT2-large fine-tuning:** 事前学習済みモデル*を使用
- **w/ LAP:** 上記に提案法を適用したモデル (テキストのみ)
- **LayoutLM** [Xu+ KDD'20]: BERTをベースにOCRで取得したレイアウト情報を組み込んだモデルを fine-tuning
- **IGBERT** [Tanaka+ ANLP'22]: 約50万件のインフォグラフィックで事前学習したBERTベースのモデル (画像 + テキスト)
- **TILT** [Powalski+ ICDAR'22]: Document AI に関するデータセット 20種 計100万件以上を使用して教師なし・あり学習で事前学習した T5 ベースのモデル (画像 + テキスト)

評価指標

- Average normalized levenshtein similarity

- 編集距離の平均値 ANLS [Biten+ ICCV'19]
 - 正解と予測との文字長を考慮した編集距離
 - OCRの誤りを考慮した正解率を緩和した指標

Results

InfographicVQA リーダーボードによる評価結果

Method	ANLS (test set)	Answer type				Evidence		Operation		
		Image span	Question span	Multiple spans	Non span	Textual	Visual object	Comparison	Arithmetic	Counting
GPT2-large	0.2456	0.2630	0.4632	0.1156	0.1521	0.2875	0.2301	0.2455	0.1790	0.1192
GPT2-large w/ LAP	0.3609	0.3593	0.4721	0.1305	0.4137	0.3783	0.3004	0.2871	0.4087	0.4063
LayoutLM [Xu+ KDD'20]	0.2720	0.3278	0.2386	0.0450	0.1371	0.3626	0.1705	0.1836	0.1559	0.1140
IG-BERT [Tanaka+ ANLP'22]	0.3854	0.4181	0.4481	0.2197	0.2849	0.5016	0.3013	0.2939	0.3564	0.2000
TILT [Powalski+ ICDAR'21]	0.6120	0.6765	0.6419	0.4391	0.3832	0.7916	0.4545	0.4801	0.4958	0.2652
Human Performance	0.9718	0.9745	0.9777	0.9335	0.9716	0.9789	0.9770	0.9712	0.9837	0.9544

GPT2-large vs. GPT2-large w/ LAP

- 提案手法を導入することで予測性能向上
 - 特に Non span, Arithmetic, Counting が向上
 - レイアウト情報の補助により、元々モデルが持つ言語情報を有効に活用して質問応答できた

提案手法はリーダーボード上で3位🏆に

- レイアウト情報を使う SoTA の LayoutLM ベースのモデルよりも予測性能の向上を確認
- 大規模なインフォグラフィックで事前学習する必要がある IGBERT と同程度の予測性能を実現
- TILT は多数の Document AI 関連のデータセットで教師あり学習をしているため、比較するのは難しい

Discussion & Future Work

レイアウト情報を導入することで言語モデルが本来有する言語・知識情報を活用可能になる

- 単に本文を抽出しても答えられない問題 Non span
- 数値を答えるような質問にも本文中の情報を元に演算可能になる (e.g., Arithmetic, Counting)

今後の展望: インフォグラフィック以外の Document AI データセットでの有効性確認

- 提案手法はレイアウト情報が重要な Document AI タスク全般に簡単に適用可能; 提案法の汎用性確認