様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １２ 日現在

機関番号：８２４０１
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７８１４

研究課題名（和文）Knowledge-Base-Grounded Language Models

研究課題名（英文）Knowledge-Base-Grounded Language Models

研究代表者
ＨＥＩＮＺＥＲＬＩＮＧ　ＢＥＮＪＡＭＩＮ（Heinzerling, Benjamin）

国立研究開発法人理化学研究所・革新知能統合研究センター・副チームリーダー

研究者番号：５０８４６４９１

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：二つの主要な成果を達成しました。最初の成果は、構造化された知識のより良い統合を可能にする言語モデル（LM）のアーキテクチャです。LMは一般的に大量のテキストデータに基づいて訓練されますが、独自の社内知識ベースなどの特定の構造化された知識を統合することが望ましい場合がよくあります。ここでは、高価な再訓練を必要とせずにそのような統合を可能にするバイエンコーダアーキテクチャを開発しました。
二つ目の成果は、LMが特定の種類の構造化知識、具体的には人の出生年や都市の人口などの数値的属性をどの程度うまく表現しているかを分析する解釈方法です。

研究成果の学術的意義や社会的意義
The first achievement provides an efficient method for integrating structured knowledge into existing language models, which allows users to adapt LMs to their specific needs without costly retraining.
The second achievement improves our understanding of how LMs, thereby increasing transparency.

研究成果の概要（英文）：This research project attained two main achievements. The first achievement is a language model (LM) architecture that enables better integration of structured knowledge. While LMs are commonly trained on large amounts of text, it is often desirable to integrate specific, structured knowledge, such as an proprietary in-house knowledge base or other knowledge that is not covered by the LM's training data. Here, we developed a bi-encoder architecture that enables such an integration without requiring costly retraining.
The second achievement is an interpretation method for analyzing how well LMs represent a specific kind of structured knowledge, namely an numeric properties such as a person's year of birth or a city's population.

研究分野：Natural Language Processing

キーワード：Language models　structured knowledge　interpretability　explainability　knowledge representation　knowledge base

１．研究開始当初の背景

At the time of application, the goal of this research project was to develop a novel language model architecture that is grounded in a structured knowledge base. The motivation for this was that humans ground text they are reading or writing in a rich set of experience, while language models lack such a grounding. For example, if a human reads a sentence like “Mount Fuji is a mountain in Japan”, she will connect the phrase “Mount Fuji” with photos of snow-capped Mount Fuji or with memories of having climbed it, she will have a mental image of a prototypical mountain which she associates with the word “mountain”, and she knows that “Japan” refers to a country in East Asia and that countries contain geographic features such as mountains. A language model, in contrast, lacks experience of the real world and instead relies only on the distributional signal of word co-occurrence patterns to produce its output. This lack of grounding often causes an ungrounded LM to produce coherent, but factually incorrect output.

２．研究の目的

The initial goal of this project was to develop a model architecture and training procedure that allows grounding a language model in a knowledge base. The underlying motivation was that this will indirectly ground the language model in the real world, since the knowledge base was created by humans whose experience is grounded in reality. Based on this indirect grounding the resulting language model should exhibit less factually incorrect output.
As large language models appeared and showed dramatic performance improvements during the second half of this project, our research focus shifted from creating a knowledge-based grounded language model architecture to the goal of understanding if and how large language models encode structured world knowledge.

３．研究の方法

During the first half of the funding period we focused on the development of the knowledge-base grounded language model architecture. Specifically, we designed an architecture that combines a knowledge-base encoder with a pretrained language model by employing cross-stitch connections and devised an efficient training procedure that avoids costly from-scratch training by using adapter modules. The resulting models showed improved performance on knowledge-intensive information extraction tasks such as relation extraction.

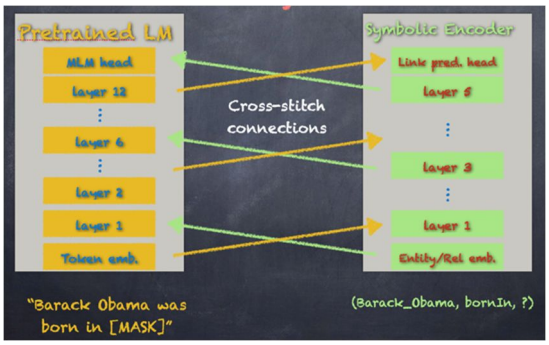

図　1 Bi-encoder architecture for combining a pretrained language model with a knowledge base encoder.

In the second half of the funding period we shifted from model developments towards developing interpretability methods for large language models with a particular on the question if and how structured is represented in the internals of such models. Concretely we developed a new method identifying and manipulating representations of numerical world knowledge in low-dimensional subspaces of a language model’s activation space (Figure 2).

４．研究成果

This research project attained two main achievements.
The first achievement is a language model (LM) architecture that enables better integration of structured knowledge. While LMs are commonly trained on large amounts of text, it is often desirable to integrate specific, structured knowledge, such as a proprietary in-house knowledge base or other knowledge that is not covered by the LM's training data. Here, we developed a bi-encoder architecture that enables such an integration without requiring costly retraining.
The second achievement is an interpretation method for analyzing how well LMs represent a specific kind of structured knowledge, namely numeric properties such as a person's year of birth or a city's population.

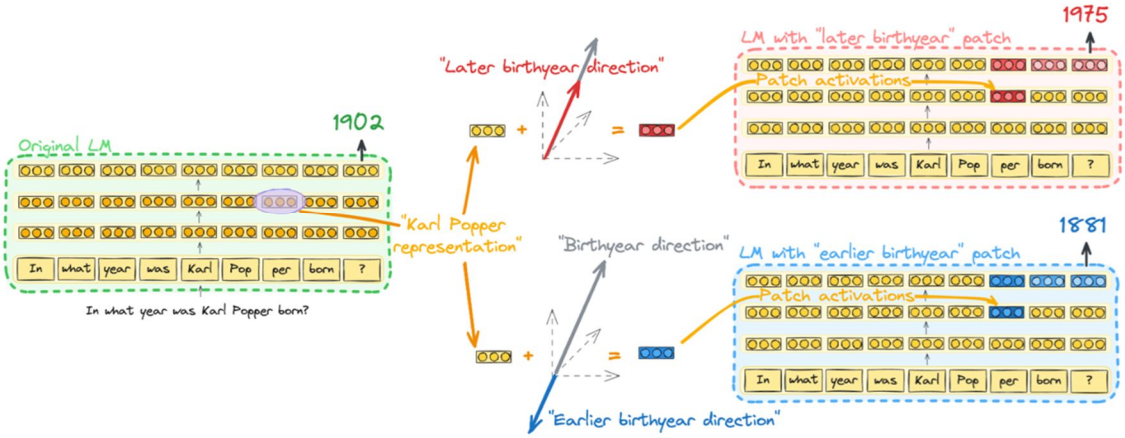

그림 2 Identification and manipulation of a “birthyear” representation in the activation space of a large language model

The first achievement provides an efficient method for integrating structured knowledge into existing language models, which allows users to adapt LMs to their specific needs without costly retraining.

The second achievement improves our understanding of how LMs, thereby increasing transparency.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　6件）

１．著者名 Qin Dai, Benjamin Heinzerling, Kentaro Inui	４．巻 1
２．論文標題 Cross-stitching Text and Knowledge Graph Encoders for Distantly Supervised Relation Extraction	５．発行年 2022年
３．雑誌名 Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing	６．最初と最後の頁 0-0
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Qin Dai, Benjamin Heinzerling, Kentaro Inui	４．巻 1
２．論文標題 Cross-stitching Text and Knowledge Graph Encoders for Distantly Supervised Relation Extraction	５．発行年 2023年
３．雑誌名 言語処理学会 第29回年次大会 発表論文集（2023年3月）	６．最初と最後の頁 0-0
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 松本 悠太, 吉川 将司, Benjamin Heinzerling, 乾 健太郎	４．巻 0
２．論文標題 ニューラル言語モデルによる一対多関係知識の記憶と操作	５．発行年 2022年
３．雑誌名 言語処理学会 第28回年次大会 発表論文集（2022年3月）	６．最初と最後の頁 556-561
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 長澤春希, Benjamin Heinzerling, 乾健太郎	４．巻 0
２．論文標題 ニューラル言語モデルによる一対多関係知識の記憶と操作	５．発行年 2022年
３．雑誌名 言語処理学会 第28回年次大会 発表論文集（2022年3月）	６．最初と最後の頁 1203-1208
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 有山知希, Benjamin Heinzerling, 乾健太郎	4．巻 0
2．論文標題 Transformer モデルのニューロンには 局所的に概念についての知識がエンコードされている	5．発行年 2022年
3．雑誌名 言語処理学会 第28回年次大会 発表論文集 (2022年3月)	6．最初と最後の頁 599-603
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Ana Brassard, Benjamin Heinzerling, Pride Kavumba and Kentaro Inui	4．巻 0
2．論文標題 COPA-SSE: Semi-structured Explanations for Commonsense Reasoning	5．発行年 2022年
3．雑誌名 Proceedings of the 13th Language Resources and Evaluation Conference	6．最初と最後の頁 (to appear)
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関