氏　　名	Han Namgi
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2272 号
学位授与の日付	2021年9月 28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Analyzing the Causal Relation Between Linguistic Knowledge and the Performance of Language Models Using Structural Equation Modeling
論 文 審 査 委 員	主　査　相澤　彰子 情報学専攻　教授 山岸　順一 情報学専攻　教授 神門　典子 情報学専攻　教授 宮尾　祐介 東京大学　大学院情報理工学系研究科　教授 鈴木　潤 東北大学　データ駆動科学・AI 教育研究センター 教授

（様式３）

博士論文の要旨

氏　　名 Han Namgi

論文題目 Analyzing the Causal Relation Between Linguistic Knowledge and the Performance of Language Models Using Structural Equation Modeling

Explaining the reason for the high performance of one system is as important as achieving high performance by using that system. Recently the language model, a vector representation of natural languages such as word2vec and BERT, has become an indispensable tool for natural language processing. While researchers have reported the state-of-the-art accuracy for a variety of downstream tasks by using language models, our understanding of this phenomenon usually depends on the observation for accuracy. However, the accuracy does not explain why one language model can obtain good accuracy and another can not. Furthermore, it is hard to find the reason for the good or bad performance of one language model for various downstream tasks from the accuracy. In other words, it indicates the lack of interpretability for language models.

Previous studies have tried to explain the quality of one language model in the aspect of encoded linguistic knowledge on that language model. However, their essential assumption, "encoded linguistic knowledge on one language model should affect the accuracy of the downstream task solved by that language model", has not been proved empirically and causally with enough samples. We present a novel framework employing the statistical method, Partial Least Squares Path Modeling (PLSPM), to explain the causal relationship between encoded linguistic knowledge and the accuracy of downstream tasks on the target language model. Our proposed framework starts from a causal diagram consisting of causal assumptions between variables, including encoded linguistic knowledge and the accuracy of downstream tasks. By validating whether the suggested causal diagram can produce similar covariance matrices with observed variables, we can examine our causal assumptions, for example, causal relationships between encoded linguistic knowledge and the accuracy of downstream tasks.

We present the usefulness of our proposed framework by following steps. First, we show that our PLSPM framework can prove the causal diagram consisting of traditional assumptions for encoded linguistic knowledge. In our PLSPM models, causal assumptions between encoded linguistic knowledge and accuracies for downstream tasks are expressed as linear regression equations. For fitting PLSPM models for our proposed causal diagrams, we prepare accuracies of one word analogy dataset measuring encoded linguistic knowledge and 20 downstream tasks solved by 600 word embedding models as observed variables. As a result, we find that our PLSPM models

can prove most causal assumptions of our causal diagrams with a variety of reliability indexes for validating the estimated PLSPM model. Comparing to previous studies, our PLSPM models provide more informative explanations for accuracies of downstream tasks involving multiple linguistic knowledge and the effect of hyperparameters on language models.

In addition, we also apply our proposed framework to more complicated language models and downstream tasks to prove that our proposed framework is also helpful in the practical setting. We conduct another PLSPM analysis involving 24 BERT models, two probing tasks, and four datasets of simple factoid question answering (SFQA), a subtask of question answering over a knowledge base. Since this task requires external resources and a modularized structured system to be solved, we select SFQA as a more complicated and practical target downstream task. The BERT-based system achieves the upper bound accuracy of SimpleQuestions, the benchmark dataset of SFQA. However, our PLSPM framework reports that this system depends on the surface and syntactic information for solving simple factoid questions without understanding semantic information. It indicates the possibility that the upper bound accuracy of existing SFQA systems for SimpleQuestions may rely on the specific characteristic of the dataset itself.

We conduct an empirical analysis involving five SFQA systems, which have reported the upper bound accuracy of SimpleQuestions, and four SFQA datasets to examine whether those systems have the robustness and transferability for SFQA. We find that all existing SFQA systems can not reach upper bound accuracies for other datasets like SimpleQuestions, and they show significantly low accuracy when changing test data. According to our analysis, the size and the upper bound accuracy of each dataset do not cause this phenomenon. We reveal that existing SFQA systems report similar problems related to semantic understanding, such as disambiguation of the entity and paraphrasing of the relation. Moreover, we suggest that the source of each dataset and the evaluation method for SFQA make existing SFQA systems depend on surface and syntactic information with the additional analysis.

In this thesis, we proposed a novel statistical framework to explain the accuracy and inner working of language models as the causal relationship with encoded linguistic knowledge. We also proved that our proposed framework could provide valuable information for understanding and resolving the encountered issue of an existing NLP system. We hope that our study can suggest a systematical and practical way to interpret the inner working of language models.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　Han Namgi

Title
論文題目　Analyzing the Causal Relation Between Linguistic Knowledge and the Performance of Language Models Using Structural Equation Modeling

出願者は、統計的言語モデルにエンコードされた言語知識と応用タスクの精度との間の因果関係を分析するために、構造方程式モデリングを適用する手法を提案し、その有用性を複数の言語モデルおよび応用タスクを用いた実験において実証した。

本学位論文は、「Analyzing the Causal Relation Between Linguistic Knowledge and the Performance of Language Models Using Structural Equation Modeling」と題し、全6章から構成されている。

第1章では、現在の自然言語処理においてword2vecに代表される単語埋め込みやBERTなどの文脈依存単語・文埋め込み等の言語モデルが重要な役割を果たしているが、これらが様々な応用タスクで示す精度の原因の分析が不十分である問題を指摘している。そして、本論文では、エンコードされた言語知識を直接測定するタスク（Intrinsic評価）の精度と応用タスクの精度との因果関係を、構造方程式モデリングを用いて分析する手法を提案している。

第2章では、背景知識として、言語モデル、構造方程式モデリングとその一手法であるPartial Least Squares Path Modeling (PLSPM)、第4章・第5章での分析対象である知識ベース質問応答について導入している。また、関連研究として言語モデルにエンコードされた言語知識を測定する Intrinsic 評価手法、言語モデルの内部動作について分析する既存手法、知識ベース質問応答の既存手法について説明し、本研究との関係を議論している。

第3章では、PLSPMを用いてIntrinsic評価の精度と応用タスクの精度の因果関係を分析する手法を提案し、その有用性を実証している。具体的には、既存研究で議論されている言語知識と応用タスクの関係を因果ダイアグラムとして定式化し、4 種類の言語知識の評価、20個の応用タスク、600個の単語埋め込みを用いてこれらの関係を分析した。その結果、言語知識と応用タスクの精度の因果関係が詳細に示され、また単語埋め込み学習のハイパーパラメータやデータセットの問題点などが明らかにされ、提案手法による分析が有用であることを実証している。

第4章では、より複雑な言語モデルおよび応用タスクの分析を行うことを目指して、知識ベース質問応答を応用タスクとし、Bidirectional Encoder Representations from Transformers (BERT)を用いた既存手法の分析を行なっている。知識ベース質問応答は、外部リソースとして知識ベースを必要とし、また複数のモジュールから構成される複合的な応用タスクである。また、BERTは現在の自然言語処理でスタンダードな言語モデルであるが、複雑なニューラルネットワークを用いて文脈依存な単語・文埋め込みを得る手法

であり、その内部動作の分析は容易でない。そこで、文埋め込みにエンコードされた言語知識の測定手法と知識ベース質問応答手法のモジュール構成に基づいて因果ダイアグラムを設計し、6 種類の言語知識の評価、3 つの質問応答データセット、24 個の BERT を用いて、PLSPM による分析を行なっている。その結果、質問応答のモジュールであるエンティティ検出や関係予測には表層的情報が主に寄与していることが示され、意味情報の寄与は棄却された。これは、既存の質問応答手法は BERT にエンコードされている意味情報を利用せず、表層的情報に頼っていることを示唆している。また、異なるデータセットでは精度や PLSPM モデルの適合度が大きく異なることから、データセットの特性が質問応答手法の性能に影響している可能性が示唆されている。

第 5 章では、第 4 章の結果に基づき、第 4 章の手法を含む 5 つの質問応答手法と、4 つのデータセットを用い、質問応答手法の頑健性と転移性、およびデータセットの特性について詳細な分析が行われている。その結果、第 4 章で示された結果は他の質問応答手法でも同様に成り立つこと、その原因はデータセットのサイズや精度上限ではなくデータにおける言語表現のバリエーションである可能性が示されている。

第 6 章では、以上の結果をまとめ、将来課題について議論を行なっている。

公開発表会では博士論文の章立てに沿って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究は十分なオリジナリティとクオリティがあるとの評価がなされた。本学位論文は、言語モデルが持つ言語知識と応用タスクの精度との関係を統計的に検証・分析する汎用的な手法を提案しており、さらに複数の言語モデルおよび応用タスクにおいて分析を行い提案手法の有用性を実証している。現在の自然言語処理モデルはブラックボックスであり内部動作の理解が困難であるという問題があるが、これに対して一つの解決策を提示したものであり、提案手法および実際の言語モデルの系統的分析は学術的な貢献が大きい。また、本学位論文の成果について、学術雑誌論文 1 件が採択され、フルペーパー査読付き国際会議論文 2 件が発表されている。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。