General Paper

Probing Simple Factoid Question Answering Based on Linguistic Knowledge

Namgi Han[†,††,†††], Hiroshi Noji[†††,†††††††], Katsuhiko Hayashi[††††], Hiroya Takamura[†††,†††††] and Yusuke Miyao[††††††]

Recent studies have indicated that existing systems for simple factoid question answering over a knowledge base are not robust for different datasets. We evaluated the ability of a pretrained language model, BERT, to perform this task on four datasets, Free917, FreebaseQA, SimpleQuestions, and WebQSP, and found that, like other existing systems, the existing BERT-based system also can not solve them robustly. To investigate the reason for this problem, we employ a statistical method, partial least squares path modeling (PLSPM), with 24 BERT models and two probing tasks, SentEval and GLUE. Our results reveal that the existing BERT-based system tends to depend on the surface and syntactic features of each dataset, and it disturbs the generality and robustness of the system performance. We also discuss the reason for this phenomenon by considering the features of each dataset and the method that was used to evaluate the simple factoid question answering task.

Key Words: *Question Answering, Language Model, Evaluation, Interpretability*

1 Introduction

Question answering over knowledge base (QAKB) has been studied as an important task in natural language processing. This task demands the prediction of entities and relations that can reach the correct answer with a knowledge base from a given question. For example, given the question, *"Which country is Albert Bolender from?"*, then the subject *Albert_Bolender* and the relation *people.person.nationality* are required to reach the correct object *United_States* in Freebase. Although various approaches have been suggested for this task, neural network-based classifier models (Yu et al. 2017; Petrochuk and Zettlemoyer 2018; Mohammed et al. 2018; Huang et al. 2019) have reported state-of-the-art accuracies recently. For example, simple factoid question

† The Graduate University for Advanced Studies
†† National Institute of Informatics
††† Artificial Intelligence Research Center, AIST
†††† Gunma University
††††† Tokyo Institute of Technology
†††††† The University of Tokyo
††††††† LeapMind Inc.

answering, a subtask in the QAKB field, is considered to be a task that has already been solved (Petrochuk and Zettlemoyer 2018). This task is a simplified version of QAKB because simple factoid questions require only one fact *(subject, relation, object)* to be solved. Many papers have reported successful accuracies with SimpleQuestions (Bordes et al. 2015), the largest benchmark dataset of simple factoid question answering. However, the high accuracy of SimpleQuestions does not mean that simple factoid question answering has been conquered.

Simple factoid questions in other datasets applicable to the QAKB task, such as Free917 (Cai and Yates 2013), WebQSP (Yih et al. 2016), and FreebaseQA (Jiang et al. 2019), are not solved, even by the systems that are successful for SimpleQuestions. Han et al. (2020b) discussed this problem by involving four datasets and systems, respectively. They revealed that existing systems such as BuboQA (Mohammed et al. 2018), HR-BiLSTM (Yu et al. 2017), KBQA-Adapter (Wu et al. 2019), and KEQA (Huang et al. 2019) cannot reach the upper bound accuracies for simple factoid questions in WebQSP and FreebaseQA. Moreover, they reported that existing systems showed a lack of transferability in the experiment across two datasets. Although their results indicate that the systems proposed in previous studies are limited in terms of general simple question answering, the effectiveness of pretrained language models for this problem has not yet been examined.

In this paper, we examine the effectiveness, transferability, and robustness of BERT by assessing its performance with respect to simple factoid question answering. Because many studies have reported successful results for various natural language processing tasks with BERT (Devlin et al. 2019; Wolf et al. 2020), we expected that BERT would be able to solve simple factoid question answering regardless of the differences among datasets. We employed an advanced version of BuboQA (Mohammed et al. 2018), which employs an LSTM and a CNN to encode a given question, by replacing all the encoders of BuboQA with BERT following Lukovnikov et al. (2019) to test our hypothesis in the same experimental setting as Han et al. (2020b). In our experiments, even though our BERT-based system attained accuracies higher than those of BuboQA on the four datasets, we also found that our BERT-based system still limited in robustness and transferability, similar to the original BuboQA.

We conducted a statistical analysis to examine the inner working of BERT to determine why BERT fails in general simple factoid question answering. Recently, several studies (Jawahar et al. 2019; Ravishankar et al. 2019; Kovaleva et al. 2019) attempted to explain the inner working of BERT using probing tasks such as SentEval (Conneau and Kiela 2018) and GLUE (Wang et al. 2019a). As a result of depending on one or a few observations on which to base their conclusions, their results may be conflicting, similar to those of previous studies that were conducted to prove

word embeddings (Schnabel et al. 2015; Chiu et al. 2016; Wang et al. 2019b; Han et al. 2020a).

Here, we propose a different approach in which this problem is defined as a statistical examination of causal relationships between the result of probing tasks and the result of simple factoid question answering. Han et al. (2020a) employed partial least squares path modeling (PLSPM) (Wold 1982) to investigate causal relationships between linguistic knowledge and NLP downstream tasks on word embeddings. Compared with other statistical methods, the advantage of PLSPM is its robustness to a smaller sample size (Tenenhaus et al. 2005a). In addition, it is easy to apply this method to analyze experimental results because it requires fewer assumptions on the observed variable (Tenenhaus et al. 2005b). Hence, we estimated PLSPM models using 24 BERT models (Turc et al. 2019) to explain the results of our BERT-based system based on the result of the probing tasks. As a result, we found that the accuracy of our BERT-based system can be causally explained by the accuracy of probing tasks on the surface and the syntactic information in our PLSPM models. This indicates that existing simple factoid question answering systems may to a large extent depend on the surface and syntactic information of the target dataset.

Our study makes the following contributions.

- Our findings showed that the system that employs pretrained language models, such as BERT, still experiences the same problems as other existing systems with respect to transferability and robustness.
- We employ PLSPM, a statistical method to probe and examine the inner working of BERT and linguistic knowledge.
- The PLSPM analysis with SentEval and GLUE revealed that the accuracies of probing tasks for semantic understanding are not causally related to the accuracies of our BERT-based system, whereas the accuracies of surface and syntactic tasks can explain the accuracies of our BERT-based system with significant path coefficients.
- Our findings suggest that the method to evaluate simple factoid question answering and the source of each dataset play an important role in this phenomenon, according to additional error analyses.

2 Background

2.1 Question answering over knowledge base

Question answering over a knowledge base is one task of semantic parsing. This task aims to find the correct answer for a given question from a target knowledge base, such as Freebase. Free917 (Cai and Yates 2013) is one of the early datasets for this task over Freebase. Free917 was

constructed by two annotators who wrote 917 questions using Freebase Commons, a subset of Freebase. However, this dataset contains only 917 questions and is too small to train a machine-learning model sufficiently. Furthermore, Berant et al. (2013) mentioned that the questions in Free917 tend to contain the label of the gold relation directly. Therefore, other datasets, such as WebQuestion (Berant et al. 2013), SimpleQuestions (Bordes et al. 2015), and FreebaseQA (Jiang et al. 2019), have been proposed on an ongoing basis.

First, WebQuestion was proposed. This dataset contains 5,810 questions that were created using Google Suggest API to ensure the naturalness of the question. Although WebQuestion succeeds in aggregating more naturally written questions than Free917, this dataset does not contain formal queries to be executed. To overcome this problem, Yih et al. (2016) suggested WebQSP, a subset of WebQuestion containing annotated SPARQL queries. SimpleQuestions is the most popular dataset for this task because of its size of over 100,000 questions. The questions in SimpleQuestions were generated by crowd workers referring to randomly sampled facts from Freebase. The approach FreebaseQA uses to aggreate questions differs from that of the above-mentioned datasets. Jiang et al. (2019) annotated 28,348 questions in TriviaQA (Joshi et al. 2017) with Freebase knowledge because Jiang et al. (2019) aimed to suggest a dataset that includes more difficult and naturally written questions such as those used in trivia quizzes.

Previously, researchers tried to solve this task by transforming questions into logical forms (Berant et al. 2013; Reddy et al. 2016; Trivedi et al. 2017). Recently, many studies have reported state-of-the-art accuracies for the above datasets by using systems based on neural network (Yu et al. 2017; Petrochuk and Zettlemoyer 2018; Mohammed et al. 2018; Huang et al. 2019). For example, BuboQA (Mohammed et al. 2018), which is well-known benchmark system for this task, consists of the following four submodules:

- entity detection, a sequence-tagging network to find the span of an entity in a given question.
- entity linking, a string-match module to link an entity in a knowledge base with the span of an entity.
- relation prediction, a sentence classifier network to predict a relation in a knowledge base from a given question.
- evidence integration, a scoring module to rerank predicted entities and relations.

Petrochuk and Zettlemoyer (2018) argued that their neural network-based model nearly solved SimpleQuestions. They examined the upper bound accuracy of SimpleQuestions and showed that their system, which employs a similar network to BuboQA, almost reached the upper bound accuracy of SimpleQuestions. However, the success of SimpleQuestions does not mean that the

QAKB task is generally successful. Han et al. (2020b) examined whether four systems, for which state-of-the-art performance was reported for SimpleQuestions, could also solve simple factoid questions in other datasets: Free917, WebQSP, and FreebaseQA. They revealed that the existing systems did not reach a level of accuracy similar to that of SimpleQuestions on other datasets. Moreover, their analysis showed that existing systems were not robust in terms of their transferability. Therefore, it is difficult to conclude that simple factoid question answering is genearlly solved.

2.2 BERT and BERTology

Contextual embeddings, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019), have become indispensable tools in natural language processing, in addition to non-contextual distributional word representations, such as word2vec (Mikolov et al. 2013) and fastText (Bojanowski et al. 2017). Despite many studies in which contextual embeddings were employed resulting in state-of-the-art performance for a variety of natural language processing tasks, their usefulness was not clearly explained. Therefore, researchers have investigated the inner working of contextual embeddings to explain their effectiveness, especially with respect to BERT.

These studies, known as BERTology (Rogers et al. 2020), usually accepted the traditional hypothesis that encoded linguistic knowledge in a language model can explain the accuracies of downstream tasks in NLP (Chiu et al. 2016). Thus, BERTology has included various analyses involving both the structural features of BERT and the linguistic knowledge encoded in BERT. The linguistic knowledge encoded in BERT has been examined in various ways, such as attention analysis (Liu et al. 2019), edge probing (Tenney et al. 2019b), and in comparison with intrinsic evaluations such as SentEval and GLUE. For example, Tenney et al. (2019a) showed that BERT encodes syntactic information, such as the part-of-speech, chucking span, and its syntactic and semantic role by edge probing analysis.

Although BERTology has succeeded in determining the type of and the way in which linguistic knowledge is encoded in BERT, they usually derived their conclusions from one or a few observations without any statistical analysis or verification. Because many language models have been proposed, observation with one or a few samples is not sufficient to make a general conclusion for BERTology. Previous studies conducted to probe word embeddings sometimes reported conflicting results (Schnabel et al. 2015; Chiu et al. 2016; Rogers et al. 2018; Wang et al. 2019b; Han et al. 2020a). Therefore, the general relationship between encoded linguistic knowledge and the performance of downstream applications is not yet clear.

2.3 Statistical analysis of word embeddings

Traditionally, encoded linguistic knowledge in the language model is believed to be helpful for solving the downstream tasks of natural language processing (Chiu et al. 2016). Before BERTology, many researchers (Schnabel et al. 2015; Chiu et al. 2016; Rogers et al. 2018; Wang et al. 2019b) have attempted to prove this intuition with distributional word representations, such as word2vec (Mikolov et al. 2013). Han et al. (2020a) argued that previous studies have two limitations. First, they only conducted a correlation analysis between the results of two tasks, for example, the accuracy of word similarity and the accuracy of POS tagging. Because a downstream task in NLP usually requires multiple linguistic knowledge to be solved, correlation analysis between the probing task and the downstream task in NLP limited ability to understand their causal relationship. Second, as we mentioned in Section 2.2, researchers sometimes reported conflicting results for the same issue because their conclusions were usually based on few observations (Schnabel et al. 2015; Chiu et al. 2016; Rogers et al. 2018; Wang et al. 2019b; Han et al. 2020a).

In the field of statistics, researchers employed structural equation modeling (SEM) (Jöreskog 1970) to prove causal assumptions between variables. In SEM, a causal diagram representing causal assumptions for the target variables is first suggested by a user. Figure 1 shows an example of a causal diagram involving probing tasks for semantic knowledge and the QAKB task. Figure 1 contains the causal assumptions listed below:

- Encoded semantic information ($y1$) affects the accuracies of probing tasks ($x1_1, x1_2, ..., x1_n$).
- The performance on the QAKB task ($y2$) affects the accuracies of entity detection ($x2_1$) and relation prediction ($x2_2$).
- Encoded semantic information ($y1$) affects the performance on the QAKB task ($y2$).

Based on a given causal diagram and the observed variables, SEM estimates the regression for-

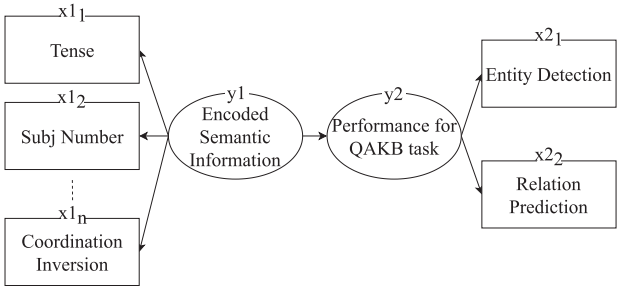


Fig. 1 Example of a causal diagram

mulas for each causal hypothesis. We refer to an estimated regression as a *structural equation*. Compared to correlation analysis, the advantage of SEM is its ability to handle multiple variables at once. Moreover, it can provide many reliable indexes for proving causal assumptions. For example, the score of $y2$ can be predicted by the following structural equation:

$$y2 = \beta_1 y1 + \zeta_1$$

where β_1 is the estimated weight, and ζ_1 is the estimated error term. The use of analysis enables us to analyze the number of latent variable scores that can be predicted by the structural equation, which the researcher has referred to as the explainability of that latent variable.

Han et al. (2020a) employed partial least squares path modeling (PLSPM) (Wold 1982), one of the methods of SEM, to investigate the relationship between the encoded linguistic knowledge in a language model and the accuracies of NLP tasks. They selected PLSPM because of its fewer requirements for observed variables and robustness with small samples. They examined their causal diagrams, which assume the causal relationship between the accuracies of probing tasks and downstream tasks, by using training algorithms, a corpus, and hyperparameters as variables. Although they proved some relationships between linguistic knowledge and the accuracy of downstream tasks in their causal diagrams, they did not consider contextual embeddings.

3 Simple Factoid Question Answering using BERT

3.1 Experimental settings

In this study, we employ the system suggested by Lukovnikov et al. (2019), an extended version of BuboQA using BERT. BuboQA, the benchmark system for simple factoid question answering (Mohammed et al. 2018), consists of four submodules: entity detection, entity linking, relation prediction, and evidence integration, as mentioned in Section 2.1. Among them, entity detection and relation prediction use a machine-learning algorithm based on a neural network, whereas the other submodules employ string matching or rule-based weight calculation. The original BuboQA employed LSTM and CNN-based encoders for entity detection and relation prediction; however, Lukovnikov et al. (2019) reported that the performance was increased by replacing LSTM and CNN-based encoders with the BERT model. Therefore, we employ the model proposed in this paper as a BERT-based simple factoid question answering system.

Note that we implemented it ourselves because the official GitHub repository of Lukovnikov et al. (2019) is no longer available. Our implementation followed the instructions in the original paper as much as possible, except for the design of the network for entity detection and relation

prediction. In their original paper, they combined two submodules into one classifier to improve the accuracy of the proposed system. However, our aim was to examine the performance of BERT compared with the original BuboQA. Therefore, we did not change the original design of BuboQA and changed only its encoder. Hereinafter, we refer to this system as BertQA to distinguish this system from the original BuboQA and that suggested by Lukovnikov et al. (2019).

The other experimental settings we used are the same as those of of Han et al. (2020b). They prepared four datasets, Free917, FreebaseQA, SimpleQuestions, and WebQSP to examine the robustness and transferability of existing systems. Because Free917, FreebaseQA, and WebQSP were not designed for simple factoid question answering, they contain questions that require multiple facts to be answered. To guarantee the same level of domain and difficulty as SimpleQuestions, Han et al. (2020b) removed all questions that require two or more facts as the answer or cannot be solved by the FB2M dataset (Bordes et al. 2014), which is the source dataset of SimpleQuestions and a subset of Freebase. We employed their filtered datasets, hereinafter F917, FBQ, SQ, and WQ, for our experiments. Table 1 provides details of the data statistics for these datasets.

In this study, we evaluated BertQA across F917, FBQ, SQ, and WQ. First, we conducted a standard evaluation on a single dataset to examine whether BertQA could reach the upper bound accuracies for each dataset, such as SimpleQuestions. In other words, this evaluation aimed to examine whether BertQA could solve simple factoid questions from the four datasets. Second, we evaluated BertQA with revised datasets that employ test splits from other datasets. For example, if we train BertQA with the train split of FBQ, then we test BertQA with the test split of SQ rather than the test split of FBQ. This evaluation aimed to examine whether BertQA can learn the ability to solve simple factoid questions that are not specific to a single dataset. Therefore, the robustness and transferability of BertQA would be evaluated in our experiments.

	Free917	FreebaseQA	SimpleQuestions	WebQSP	F917	FBQ	SQ	WQ
Train	512	20,358	75,910	2,478	—	10,427	75,895	1,292
Valid	129	3,994	10,845	620	—	2,048	10,843	323
Test	276	3,996	21,687	1,639	347	2,102	21,680	861

Table 1 Data statistics for datasets. Note that we do not use F917 as a training dataset, because of its small size.

3.2 Results

Table 2 lists the end-to-end top-1 accuracies of BertQA for F917, FBQ, SQ, and WQ. Our findings indicated that BertQA achieves higher accuracies than BuboQA, except for only one case. However, BertQA does not reach the upper bound accuracy[1] of FBQ and WQ, similar to SQ. Han et al. (2020b) reported the upper bound accuracies of FBQ (84%) and WQ (87%), which are similar to the upper bound accuracy of SimpleQuestions, 83% (Petrochuk and Zettlemoyer 2018). In Table 2, the accuracies of BertQA for FBQ and WQ were much lower than the upper bound accuracies of FBQ and WQ. This indicates that BertQA has difficulty solving simple factoid questions in FBQ and WQ.

Table 2 indicates that BertQA becomes weaker when the training data and the test data are different, even though the level of domain and difficulty of the datasets are the same. This indicates that BertQA shows a lack of transferability. To examine the reason for the lower transferability, we examined error cases when BertQA, which was trained with the train split of WQ, tried to solve the validation split of WQ. We only considered the SQ and WQ cases because Han et al. (2020b) reported that FBQ contains unfaithful questions. For example, we found that FBQ annotated a gold subject *Africa* and a gold relation *location.contains* for a question *What is the highest volcano in Africa?*. This question should be answered by multiple facts consisting of a fact for volcanos in Africa and the fact that it restricts the answer to the highest volcano.

Train	Test	BertQA	Comparison with BuboQA	Comparison with upper bound	Comparison with same train test
FBQ	F917	23.63	+6.34	—	—
	FBQ	42.29	+4.04	−41.71	—
	SQ	32.07	+8.30	−50.93	−42.00
	WQ	38.88	+9.78	−48.12	−25.96
SQ	F917	47.84	+6.92	—	—
	FBQ	24.41	+4.33	−59.59	−17.88
	SQ	74.07	−0.74	−08.93	—
	WQ	44.59	+2.80	−42.41	−20.25
WQ	F917	14.70	+2.02	—	—
	FBQ	10.28	+2.34	−73.72	−32.01
	SQ	19.61	+3.15	−63.39	−54.46
	WQ	64.84	+3.61	−22.16	—

Table 2 Results of BertQA and BuboQA for the datasets. The reported value for the upper bound accuracies of each dataset was the same as that employed in previous studies (Petrochuk and Zettlemoyer 2018; Han et al. 2020b).

[1] Note that we do not consider the upper bound accuracy of F917, which consists of only the test split in our experiment because of its small size.

Label	Description	# of questions
relnotfound	target relation is not found in train data	7
wrongent	predicted entity is wrong	5
wrongrel	predicted relation is wrong	21
ambient	predicted entity can reach answer, but is not the same as gold entity	5
ambirel	predicted relation can reach answer, but is not the same as gold relation	24
null	no predicted entity or relation	13
total		75

Table 3 Labeling for questions on WQ validation split, which fails when the training data changes from WQ to SQ.

However, FreebaseQA annotated this question as a simple factoid question.

Because it indicates that the problems BertQA experiences when solving FBQ may be caused by FBQ itself, we excluded FBQ from further analyses in this study.

Table 3 indicates that approximately 70% of questions failed with resprect to relation prediction. This problem was also reported by Han et al. (2020b), who argued that the errors in relation prediction were caused by the discrepancies in question styles across the datasets. For example, the *people.person.profession* relation is mainly related to the *Who is ...?* question in WQ. However, the same relation tends to be related to *What is ... 's profession?* question in SQ. According to Han et al. (2020b), the four state-of-the-art systems for SimpleQuestions have the same problem with relation prediction, regardless of the difference in the structure and the approach followed by those systems. Considering that our experiment with BertQA experiences the same problem as Han et al. (2020b), a pretrained language model may not be the solution for general simple factoid question answering.

4 Statistical Analysis of Simple Factoid Question Answering with BERT

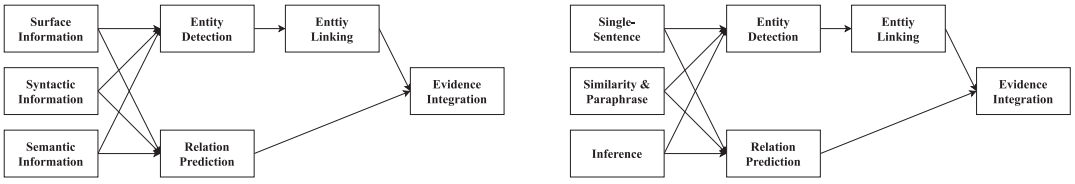
In this section, we present a statistical analysis of the results of BertQA to understand why BertQA still has the same problem as other existing systems. Han et al. (2020b) attempted to explain the results they obtained for simple factoid question answering systems, but their analyses were limited to ad-hoc error analysis. This prompted us to employ PLSPM analysis, which is a priori way to investigate the inner working of existing systems.

4.1 Causal diagram and data preparation

In this study, we aimed to investigate the inner working of BertQA when solving simple factoid question answering. As explained in Section 2.3, causal assumptions for the target variables should be expressed as a causal diagram as motivation to employ PLSPM. We drew inspiration for our causal diagram from previous studies, which assumed that the accuracies of probing tasks can explain the accuracies of NLP tasks (Chiu et al. 2016) on the analysis of word embedding models. Following the traditional assumption, we suggest the use of the causal diagrams shown in Figure 2, which consist of causal hypotheses among linguistic knowledge and simple factoid question answering. Our causal diagrams assume that the same language model solves all probing tasks and simple factoid question answering. Because the probing task was designed to examine encoded linguistic knowledge in the language model (Conneau and Kiela 2018), we only employed BertQA, which only consists of BERT and one dense decoder, in this study.

Our causal diagrams can be divided into two parts: a part for probing tasks as the cause and the other part for submodules in the BertQA system as the effect. Referring to previous studies (Jawahar et al. 2019; Kovaleva et al. 2019; Hao et al. 2019; Schneider et al. 2020), we employed SentEval (Conneau and Kiela 2018) and GLUE (Wang et al. 2019a) as probing tasks to examine the accuracy of the encoded linguistic knowledge in BERT. Then, we divided BertQA into its submodules, including entity detection, entity linking, relation prediction, and evidence integration, following the original paper of BuboQA. Because entity linking and evidence integration do not use the BERT-based machine-learning system, we set causal hypotheses only between the probing tasks and two submodules of BertQA, including entity detection and relation prediction. Therefore, our PLSPM models based on the diagrams shown in Figure 2 aim to estimate structural equations for the following hypotheses:

- The accuracies of probing tasks, SentEval and GLUE, can explain the accuracy of entity detection and the accuracy of relation prediction.



(a) Causal diagram between SentEval and BertQA (b) Causal diagram between GLUE and BertQA

Fig. 2 Causal diagrams

- The accuracy of entity detection can explain the accuracy of entity linking.
- The accuracy of entity linking and the accuracy of relation prediction can explain the accuracy of evidence integration.

We employed 24 BERT models from Turc et al. (2019) to prepare sample data for our causal diagrams. These BERT models share their standard training recipe with BERT-Base (Devlin et al. 2019), only differing in the number of layers (2, 4, 6, 8, 10, 12) and hidden embedding sizes (128, 256, 512, 768). Even though PLSPM allows a smaller sample size compared with other structural equation modeling methods (Tenenhaus et al. 2005b; Sanchez 2013), we did not combine SentEval and GLUE into one PLSPM model to decrease the number of parameters of PLSPM models. Furthermore, we estimated PLSPM models for FBQ, SQ, and WQ, respectively; that is, we prepared six PLSPM models in total. Hereinafter, we use the following naming convention for a PLSPM model estimated with the accuracies of the specific probing dataset and BertQA for a simple factoid question answering dataset: `(the name of the probing dataset)-(the name of the dataset for simple factoid question answering)`. For example, *SentEval-FBQ* indicates that this PLSPM model is estimated with the accuracies of SentEval and the accuracies of BertQA for the FBQ dataset.

For the observed variables in our PLSPM models, we prepared the results of SentEval, GLUE, and submodules of BertQA with 24 BERT models. Table 4 provides the details of the tasks we employed. Note that we did not conduct any preprocessing for the observed variables except for normalization, whereas SentEval and GLUE used many kinds of performance indicators, such as Top-n accuracy, F1-score, and Matthews correlation coefficient. For our causal diagram, we need

Dataset-Category as latent variable	Tasks as observed variables
SentEval-Surface Information (SUR)	Length, Word Content
SentEval-Syntactic Information (SYN)	Tree Depth, Bigram Shift, Top-constituent
SentEval-Semantic Information (SEM)	Tense, Subj Number, Obj Number, Odd Man Out, Coordination Inversion
GLUE-Single-Sentence (SS)	CoLA, SST-2
GLUE-Similarity and Paraphrase (SP)	MRPC, STS-B, QQP
GLUE-Inference (IF)	MNLI, QNLI, RTE, WNLI
SFQA-Entity Detection (ED)	Entity Detection
SFQA-Entity Linking (EL)	Entity Linking
SFQA-Relation Prediction (RP)	Relation Prediction
SFQA-Evidence Integration (EI)	Evidence Integration

Table 4 List of tasks used for PLSPM models. Gray tasks were not used in our experiments because of the low correlation coefficients.

to define the original indicator for measuring the performance of each task. Furthermore, we also structured our employed tasks according to their original paper to obtain the composite latent variables, except for QQP and WNLI in the GLUE dataset. These tasks have low correlation coefficients with other tasks in the same category, which negatively affects the reliability of PLSPM models. Our code will be shared in the GitHub repository for reproducibility.

4.2 Results

First, we examine the extent to which the accuracy of SentEval can predict the accuracy of BertQA by PLSPM analysis. Table 5 lists the path coefficients of each path in SentEval-FBQ, SentEval-SQ, and SentEval-WQ. When the PLSPM model is interpreted in the statistical field, the path coefficient is considered as the explainability of the target path. We do not list paths for which the p-value is larger than 0.05. Note that it is considered to be a meaningful index even though the path coefficient is a negative value because a meaningless path would be rejected on the basis of the p-value. As a result, we can conclude that the accuracies of SentEval can meaningfully explain the accuracies of BertQA, except for the semantic tasks of SentEval. This indicates that BertQA cannot overcome the discrepancy between different datasets because the inconsistency in the distribution of questions between different datasets is related to surface and syntactic knowledge.

We find it to be more difficult to interpret the result of PLSPM models using the GLUE dataset as is clear from Table 6. Only the accuracies of the inference tasks can explain the accuracy of entity detection with the p-value < 0.05 among GLUE tasks, yet they also report negative coefficients for explaining the accuracies of relation prediction. In this regard, the accuracies of single-sentence tasks, such as CoLA, mainly explain the relation prediction. One

FBQ	Surface	Syntactic	Semantic
Entity Detection	−0.544	+1.060	—
Relation Prediction	+0.327	+0.405	—
SQ	**Surface**	**Syntactic**	**Semantic**
Entity Detection	−0.674	+1.190	—
Relation Prediction	+0.462	—	—
WQ	**Surface**	**Syntactic**	**Semantic**
Entity Detection	−0.590	+0.976	—
Relation Prediction	+0.418	—	—

Table 5 Path coefficient for PLSPM models with SentEval. If the p-value of the path equation is higher than 0.05, we reject this path.

interesting point is that the accuracies of the similarity and paraphrase tasks are rejected to explain the accuracy of BertQA with a p-value > 0.05. These datasets of the similarity and paraphrase tasks, such as MRPC, require an understanding of the semantic knowledge of the given sentences to solve the questions. This means that encoded semantic knowledge and the ability to understand the given sentences are not very helpful in explaining the accuracies of BertQA, even in our PLSPM model with the GLUE dataset.

In terms of the overall results of our PLSPM models, we found another problem regarding the discrepancy between each simple factoid question answering dataset. Table 7 lists the goodness-of-fit indexes of each PLSPM model. The goodness-of-fit value of the PLSPM model indicates the extent to which this model can explain the observed and latent variables. Although all PLSPM models share the same causal diagram, the PLSPM models for the WQ dataset reported lower goodness-of-fit indices than the PLSPM models for other datasets. In particular, SentEval-WQ shows a lower goodness-of-fit index than 0.6, which indicates that this model is not well explained by a given causal diagram. This result implies that WQ is solved with different encoded linguistic knowledge compared with FBQ and SQ.

FBQ	Single-Sentence	Similarity & Paraphrase	Inference
Entity Detection	—	—	+0.756
Relation Prediction	+1.460	—	−0.780
SQ	Single-Sentence	Similarity & Paraphrase	Inference
Entity Detection	—	—	+1.090
Relation Prediction	+1.340	—	−0.918
WQ	Single-Sentence	Similarity & Paraphrase	Inference
Entity Detection	—	—	+1.910
Relation Prediction	+1.400	—	—

Table 6 Path coefficient for PLSPM models with GLUE. If the p-value of the path equation is higher than 0.05, we reject this path.

	Goodness-of-Fit
SentEval-FBQ	0.6826
SentEval-SQ	0.7448
SentEval-WQ	0.5533
GLUE-FBQ	0.8218
GLUE-SQ	0.8783
GLUE-WQ	0.6769

Table 7 Goodness-of-Fit index for each of the PLSPM models.

Another issue is the discrepancy in the GoF values between the PLSPM models using SentEval and the PLSPM models using GLUE. As in Table 7, we find that the PLSPM models using GLUE have higher GoF values than those using SentEval. This indicates that the linguistic knowledge measured by SentEval has lower explainability than that measured by GLUE, following our causal diagrams in Figure 2. We suppose two reasons for this phenomenon. First, Conneau and Kiela (2018) reported that sometimes the accuracy of their probing task does not show any correlation with the accuracy of downstream tasks. For example, the tree depth and bigram shift tasks only correlate with one downstream task among the 17 downstream tasks. This means that the probing tasks in SentEval are not sufficiently robust to explain the accuracy of downstream tasks. Second, SentEval was proposed for sentence encoders such as SkipThought (Kiros et al. 2015) and InferSent (Conneau et al. 2017) before the proposal of contextual embeddings. We suppose that this is why the probing task in SentEval has lower explainability for the BERT-based model.

5 Discussion

5.1 BertQA and semantic understanding

The results in Table 5 show that BertQA largely depends on the encoded surface and syntactic knowledge. This also means that the surface and syntactic features of each dataset affect the accuracy of the BertQA model. On the other hand, the accuracy of semantic information tasks cannot explain the accuracies of BertQA. As mentioned in Section 3.2, semantic understanding is required to solve ambiguous questions among datasets, such as *people.person.profession*. Therefore, the lack of semantic understanding in BertQA is an important factor responsible for the failure of general simple factoid question answering.

The method used to evaluate the simple factoid question answering task, namely matching the predicted subject and relation with gold data (Bordes et al. 2015), is one reason for this problem. Particularly in SimpleQuestions, the subject and the relation can be extracted from the question without semantic understanding, because the questions usually contain labels of the subject and the relation (Serban et al. 2016). This means that the evaluation method compels existing QA models to concentrate on surface and syntactic features. Furthermore, the possibility of multiple correct facts for the given questions could be another problem with this evaluation method. For example, when BertQA predicted *people.person.profession* for a given question then it was able to find the correct answer in Freebase, but the traditional evaluation method may reject this prediction if the gold relation in the dataset is *common.topic.notable_type*. Hereinafter, we refer to the traditional evaluation method as the *matching accuracy* because the criteria of

this evaluation are based on matching the correct subject and relation from a given question.

To examine the effect of the evaluation method on the accuracy of BertQA, we conducted additional experiments. We employed an older evaluation method (Berant et al. 2013; Berant and Liang 2014), which considers an object, the answer to a given question in the QAKB task, to overcome the limitation of the matching accuracy. Evidence integration, a submodule of BertQA, combines the predicted subjects and predicted relations for a given input question, and evaluates its result with the gold fact given by the dataset. Here, our employed evaluation method compares the predicted object, which is derived from a predicted subject and relation, with the gold object in the dataset. Moreover, we extended this method to entity linking and relation prediction. In particular, we aggregated all facts from FB2M to automatically examine whether the predicted result of each submodule can reach the gold object. We refer to this employed evaluation method as the *reachability accuracy* because the criteria of this evaluation are based on the reachability of the correct answer in a knowledge base.

Table 8 compares the matching and reachability accuracy for the experimental results. BertQA still does not reach the upper bound accuracies of each dataset, even when evaluating the reachability accuracy. However, BertQA achieves higher accuracies on average with the reachability accuracy. This means that many entities and relations, which can reach the correct answer in a knowledge base, were scored as incorrect predictions with the previous evaluation method.

Train	Test	Matching Accuracy	Reachability Accuracy
FBQ	F917	23.63	27.38
	FBQ	42.29	51.47
	SQ	32.07	36.37
	WQ	38.88	40.75
SQ	F917	47.84	51.30
	FBQ	24.41	32.57
	SQ	74.07	78.05
	WQ	44.59	44.70
WQ	F917	14.70	15.85
	FBQ	10.28	12.89
	SQ	19.61	21.97
	WQ	64.84	62.40
Average		36.43	39.64

Table 8 Results of BertQA for QAKB datasets. The matching accuracy was calculated by checking whether the predicted subject and relation were the same as the gold data. The reachability accuracy was calculated by checking whether the predicted subject and relation can reach the gold object.

	Goodness-of-Fit using matching accuracy	Goodness-of-Fit using reachability accuracy
SentEval-FBQ	0.6826	0.7095
SentEval-SQ	0.7448	0.7323
SentEval-WQ	0.5533	0.5891
GLUE-FBQ	0.8218	0.8673
GLUE-SQ	0.8783	0.8721
GLUE-WQ	0.6769	0.7127

Table 9 Comparison of Goodness-of-Fit indexes between evaluation methods.

This indicates that the previous evaluation method does not consider paraphrasing or synonyms, strongly related to semantic understanding.

We also estimated the PLSPM models with the reachability accuracies of BertQA. Note that we did not change our causal models at all, because we only employed the reachability accuracies of BertQA as observed variables. The new PLSPM models still rejected the structural equations between the probing tasks for semantic information and BertQA. However, the new PLSPM models reported a higher Goodness-of-Fit value for all datasets on average than the PLSPM models with matching accuracies, as indicated in Table 9. In particular, in Table 10, the difference in the R^2 value of latent variables is generally larger for WQ than for SQ. Because it is related to the features of WebQuestions, we continue this discussion in the next section.

5.2 Special characteristics of WQ

As we reported in Section 4.2, the Goodness-of-Fit values of the PLSPM models for WQ are lower than those of the PLSPM models for other datasets. This indicates that the same causal diagram, which assumes that encoded linguistic knowledge can explain the accuracies of BertQA, does not match for WQ, unlike other datasets. Table 10 contains the R^2 value of each variable in our PLSPM models. R^2 denotes the explanatory power of each variable. For example, the R^2 value of 0.787 for ED means that approximately 79% of the ED variables can be explained by the PLSPM model. We found that the observed variables of entity detection (ED) and relation prediction (RP) for WQ are not explained more satisfactorily by linguistic knowledge than ED and RP for other datasets. Furthermore, the result obtained for entity linking (EL) for WQ is not explained by the accuracy of entity detection for WQ, unlike the result obtained for other datasets. These results indicate that WQ is an especially difficult dataset for predicting the correct entity from a given question using the existing QA model. Therefore, we conducted a qualitative analysis with question sentences of WQ to determine the reason for this phenomenon.

We investigated the error cases of BertQA for the validation split of WQ to examine why

Model	Variable	When using matching accuracy	When using reachability accuracy
SentEval-FBQ	ED	0.787	0.797
	EL	0.345	0.758
	RP	0.880	0.713
	EI	0.988	0.979
GLUE-FBQ	ED	0.890	0.887
	EL	0.345	0.758
	RP	0.868	0.816
	EI	0.988	0.979
SentEval-SQ	ED	0.820	0.820
	EL	0.935	0.836
	RP	0.851	0.841
	EI	0.961	0.953
GLUE-SQ	ED	0.884	0.885
	EL	0.935	0.836
	RP	0.748	0.805
	EI	0.961	0.953
SentEval-WQ	ED	0.376	0.406
	EL	0.106	0.041
	RP	0.738	0.875
	EI	0.754	0.923
GLUE-WQ	ED	0.578	0.581
	EL	0.106	0.041
	RP	0.660	0.778
	EI	0.754	0.923

Table 10 R^2 (explanatory power) for each variable in our PLSPM models. A higher value indicates a higher explainablilty for the target variable. ED is entity detection, EL is entity linking, RP is relation prediction, and EI is the end-to-end accuracy from evidence integration.

these problems occur. We manually labeled 20 error questions, which were evaluated as being correct in entity detection but as incorrect in entity linking, as in Table 11. As a result, we find that two main problems occur when linking the entity in given questions. First, Freebase links too many entities with a single entity label. For example, when we find an entity with the label *mexico* in the preprocessed index by BuboQA, we obtain 2,830 results. Because the entity linking of BuboQA and BertQA does not have any scoring process to sort ambiguous results, BertQA sometimes cannot find the correct entity for a given question within the top-n results. The second problem is that the string label for the entity in Freebase and the written string for the subject in WQ are not identical. For example, the given question in WQ has the phrase, *communist party of china*, but the label of the entity for this phrase in Freebase is *Chinese communist party*.

Therefore, the characteristics of WQ are the reason why entity detection and linking of WQ are not explained well by our PLSPM models.

For relation prediction, we find that sometimes the same relation is written with different patterns by annotators on each dataset. We examined the relation *people.person.profession* as the sample relation to investigate the difference in writing patterns between datasets. If the term *profession* appears directly in the question, we label it as *directly specify relation*. If a term *job, work, ...* that is similar to *profession* appears in the question, we use the *indirectly specify relation label*, and if there is no term similar to *profession* in the question, *paraphrasing* is used for annotation. Table 12 contains the result of our annotation on the validation split of SQ and WQ. In SQ, approximately 80% of the questions for *people.person.profession* contain a term that indicates the relation either directly or indirectly, whereas in WQ only 13% of the questions contain such a term. This result indicates that the relation prediction of WQ demands more complex linguistic knowledge than surface and syntactic understanding, unlike SQ. Therefore, the relation prediction of WQ is not explained more satisfactorily than that of SQ in our PLSPM

Label	Description	Example	Number
DiffString	Entity label in Freebase and written strings are different	who created the chinese communist party, communist party of china	11
WrongKB	Freebase has too many or no entities for one label	what continent is mexico located on, mexico	6
ErrorDetection	Entity detection is evaluated as correct, but actually fails	what is the northeast of the united states, united states	3
Total			20

Table 11 Twenty error cases in entity linking for WQ.

Dataset	Label	Example	Number
SQ-Validation	directly specify relation	name the **profession** of peter heller.	84
	indirectly specify relation	what **job** does jamie hewlett have?	14
	paraphrasing	what does dan osborn do for a living?	25
	total		123
WQ-Validation	directly specify relation	—	—
	indirectly specify relation	what **job** does bill rancic have?	3
	paraphrasing	who is henry david thoreau?	20
	total		23

Table 12 Question patterns for the relation *people.person.profession* in SQ and WQ. Number means the number of questions requiring *people.person.profession* as the gold relation.

models.

As discussed above, WQ has different characteristics for entities and relations compared with other datasets, especially SQ. According to our analysis, entity linking for WQ requires the ability to disambiguate too many candidate entities for the same label in Freebase. Furthermore, relation prediction for WQ involves semantic understanding to predict the gold relation because the questions in WQ tend to contain a paraphrased term or synonym of the label for the gold relation. Han et al. (2020b) reported that BuboQA and KEQA, other state-of-the-art systems for SimpleQuestions, have the same difficulty with entity linking and relation prediction for WQ as BertQA. This indicates that WQ is a more challenging dataset than SQ, even for other state-of-the-art systems.

One main reason for this phenomenon is the difference in the method that is used to generate questions in the dataset. For WebQuestions, the source of WQ, the questions were generated by Google Suggest API as naturally written user queries (Berant et al. 2013). On the other hand, the question in SimpleQuestions, the source of SQ, were written artificially by crowd workers based on the suggested fact (Bordes et al. 2015). Therefore, the difference in the method that was used to create each dataset is the reason for the difference in distribution among the datasets. Moreover, this is also the reason why each dataset is solved using different linguistic knowledge in our PLSPM models.

6 Conclusion

In this paper, we examined whether the pretrained language model, BERT, could robustly solve simple factoid question answering. BERT-based models proved to be robust in other natural language processing areas (Talmor and Berant 2019). However, we found that BertQA failed to solve various tasks successfully, as in previous systems. We conducted PLSPM analysis to investigate why BertQA could not overcome the discrepancy in the distribution among datasets. As a result, our experiments revealed that even BERT depends on the surface and syntactic features of each dataset, rather than the semantic understanding required for general simple factoid question answering. This indicates that a pretrained language model is not sufficient to generalize the distribution discrepancy among existing datasets. In addition, we discuss the source of each dataset and the particular method that was used to evaluate simple factoid question answering, which are important to understand why even BERT depends on the surface and syntactic features.

Han et al. (2020b) mentioned other possible approaches to general simple factoid question

answering, such as improving the quality of the dataset and distributionally robust optimization (Delage and Ye 2010; Oren et al. 2019) for generalizing biased datasets. In addition, new datasets have been proposed that consider the distributions and difficulties of questions (Gu et al. 2021). Although these studies yielded promising results, we additionally suggest that it would also be necessary to reconsider the evaluation method for simple factoid question answering. For example, changing an objective function from the subject and relation to an object may improve the semantic understanding of the QA system for the given questions. In future work, we hope to suggest a more robust system for simple factoid question answering, based on the findings of this study and those of other researchers.

Acknowledgement

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). For experiments, computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). "Semantic Parsing on Freebase from Question-Answer Pairs." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Berant, J. and Liang, P. (2014). "Semantic Parsing via Paraphrasing." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics*, **5**, pp. 135–146.

Bordes, A., Chopra, S., and Weston, J. (2014). "Question Answering with Subgraph Embeddings." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 615–620, Doha, Qatar. Association for Computational Linguistics.

Bordes, A., Usunier, N., Chopra, S., and Weston, J. (2015). "Large-scale Simple Question Answering with Memory Networks." *ArXiv*, **abs/1506.02075**.

Cai, Q. and Yates, A. (2013). "Large-scale Semantic Parsing via Schema Matching and Lexicon Extension." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 423–433, Sofia, Bulgaria. Association for Computational Linguistics.

Chiu, B., Korhonen, A., and Pyysalo, S. (2016). "Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance." In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 1–6.

Conneau, A. and Kiela, D. (2018). "Senteval: An Evaluation Toolkit for Universal Sentence Representations." *arXiv preprint arXiv:1803.05449*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Delage, E. and Ye, Y. (2010). "Distributionally Robust Optimization under Moment Uncertainty with Application to Data-driven Problems." *Operations Research*, **58** (3), pp. 595–612.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gu, Y., Kase, S., Vanni, M., Sadler, B., Liang, P., Yan, X., and Su, Y. (2021). "Beyond IID: Three Levels of Generalization for Question Answering on Knowledge Bases." In *Proceedings of the Web Conference 2021*, pp. 3477–3488.

Han, N., Hayashi, K., and Miyao, Y. (2020a). "Analyzing Word Embedding Through Structural Equation Modeling." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1823–1832, Marseille, France. European Language Resources Association.

Han, N., Topic, G., Noji, H., Takamura, H., and Miyao, Y. (2020b). "An Empirical Analysis of Existing Systems and Datasets toward General Simple Question Answering." In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5321–5334, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hao, Y., Dong, L., Wei, F., and Xu, K. (2019). "Visualizing and Understanding the Effectiveness of BERT." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4143–4152, Hong Kong, China. Association for Computational Lin-

guistics.

Huang, X., Zhang, J., Li, D., and Li, P. (2019). "Knowledge Graph Embedding Based Question Answering." In *Proceedings of the 20th ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 105–113, New York, NY, USA. ACM.

Jawahar, G., Sagot, B., and Seddah, D. (2019). "What Does BERT Learn about the Structure of Language?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jiang, K., Wu, D., and Jiang, H. (2019). "FreebaseQA: A New Factoid QA Data Set Matching Trivia-Style Question-Answer Pairs with Freebase." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Jöreskog, K. G. (1970). "A General Method for Analysis of Covariance Structures." *Biometrika*, **57** (2), pp. 239–251.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). "Skip-Thought Vectors." *arXiv preprint arXiv:1506.06726*.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). "Revealing the Dark Secrets of BERT." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). "Linguistic Knowledge and Transferability of Contextual Representations." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Lukovnikov, D., Fischer, A., and Lehmann, J. (2019). "Pretrained Transformers for Simple Question Answering over Knowledge Graphs." In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I*, pp. 470–486.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). "Linguistic Regularities in Continuous Space word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751.

Mohammed, S., Shi, P., and Lin, J. (2018). "Strong Baselines for Simple Question Answering over Knowledge Graphs with and without Neural Networks." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

Oren, Y., Sagawa, S., Hashimoto, T., and Liang, P. (2019). "Distributionally Robust Language Modeling." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4227–4237, Hong Kong, China. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). "Deep Contextualized Word Representations." *arXiv preprint arXiv:1802.05365*.

Petrochuk, M. and Zettlemoyer, L. (2018). "SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 554–558, Brussels, Belgium. Association for Computational Linguistics.

Ravishankar, V., Gökırmak, M., Øvrelid, L., and Velldal, E. (2019). "Multilingual Probing of Deep Pre-Trained Contextual Encoders." In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pp. 37–47, Turku, Finland. Linköping University Electronic Press.

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., and Lapata, M. (2016). "Transforming Dependency Structures to Logical Forms for Semantic Parsing." *Transactions of the Association for Computational Linguistics*, **4**, pp. 127–140.

Rogers, A., Hosur Ananthakrishna, S., and Rumshisky, A. (2018). "What's in Your Embedding, And How It Predicts Task Performance." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). "A Primer in Bertology: What We Know about How Bert Works." *arXiv preprint arXiv:2002.12327*.

Sanchez, G. (2013). "PLS Path Modeling with R." *Berkeley: Trowchez Editions*, **383**, p. 2013.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). "Evaluation Methods for Unsu-

pervised Word Embeddings." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 298–307.

Schneider, R., Oberhauser, T., Grundmann, P., Gers, F. A., Loeser, A., and Staab, S. (2020). "Is Language Modeling Enough? Evaluating Effective Embedding Combinations." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4739–4748, Marseille, France. European Language Resources Association.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., and Bengio, Y. (2016). "Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 588–598, Berlin, Germany. Association for Computational Linguistics.

Talmor, A. and Berant, J. (2019). "MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4911–4921, Florence, Italy. Association for Computational Linguistics.

Tenenhaus, M., Pages, J., Ambroisine, L., and Guinot, C. (2005a). "PLS Methodology to Study Relationships between Hedonic Judgements and Product Characteristics." *Food quality and preference*, **16** (4), pp. 315–325.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005b). "PLS Path Modeling." *Computational Statistics & Data Analysis*, **48** (1), pp. 159–205.

Tenney, I., Das, D., and Pavlick, E. (2019a). "BERT Rediscovers the Classical NLP Pipeline." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., et al. (2019b). "What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations." In *7th International Conference on Learning Representations, ICLR 2019*.

Trivedi, P., Maheshwari, G., Dubey, M., and Lehmann, J. (2017). "LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs." In d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., and Heflin, J. (Eds.), *The Semantic Web – ISWC 2017*, pp. 210–218, Cham. Springer International Publishing.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Well-read Students Learn Better: On the Importance of Pre-training Compact Models." *arXiv preprint arXiv:1908.08962*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). "GLUE:

A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019b). "Evaluating Word Embedding Models: Methods and Experimental Results." *APSIPA Transactions on Signal and Information Processing*, **8**, p. e19.

Wold, H. (1982). "Soft Modeling: The Basic Design and Some Extensions." *Systems Under Indirect Observation*, **2**, p. 343.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online. Association for Computational Linguistics.

Wu, P., Huang, S., Weng, R., Zheng, Z., Zhang, J., Yan, X., and Chen, J. (2019). "Learning Representation Mapping for Relation Detection in Knowledge Base Question Answering." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6130–6139, Florence, Italy. Association for Computational Linguistics.

Yih, W.-t., Richardson, M., Meek, C., Chang, M.-W., and Suh, J. (2016). "The Value of Semantic Parse Labeling for Knowledge Base Question Answering." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206, Berlin, Germany. Association for Computational Linguistics.

Yu, M., Yin, W., Hasan, K. S., dos Santos, C., Xiang, B., and Zhou, B. (2017). "Improved Neural Relation Detection for Knowledge Base Question Answering." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 571–581, Vancouver, Canada. Association for Computational Linguistics.

Namgi Han: received his Ph.D. from The Graduate University for Advanced Studies, SOKENDAI, in 2021. He is currently a post-doctoral fellow at Ulsan National Institute of Science and Technology. His research interests include natural language processing and neural network interpretability.

Hiroshi Noji: received his Ph.D. from the Graduate University for Advanced Studies and was an assistant professor at Nara Institute of Science and Tech-

nology. He is currently a machine learning engineer at LeapMind Inc. and a visiting researcher at AI Research Center of Advanced Industrial Science and Technology. His research interests include computational linguistics, in particular syntactic parsing.

Katsuhiko Hayashi: received PhD in Engineering from the graduate school of Information Science at the Nara Institute of Science and Technology in 2013. He is currently an associate professor in Gunma University and a visiting researcher in the Riken Center for Advanced Intelligence Project. His research interests include natural language processing and statistical relational learning.

Hiroya Takamura: received his Ph.D. from Nara Institute of Science and Technology. He worked as a professor at Tokyo Institute of Technology, and currently is a research team leader at AI Research Center of Advanced Industrial Science and Technology. His current research interests include natural language processing.

Yusuke Miyao: received his Ph.D. from the University of Tokyo in 2006. He has been Assistant Professor at the University of Tokyo from 2001 to 2010, Associate Professor at National Institute of Informatics since 2010, and Professor at the University of Tokyo from 2018.

(Received March 1, 2021)
(Revised May 31, 2021)
(Accepted July 5, 2021)