Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1823–1832
Marseille, 11–16 May 2020

Analyzing Word Embedding Through Structural Equation Modeling

Namgi Han[12], **Katsuhiko Hayashi**[3], **Yusuke Miyao**[3]
[1]The Graduate University for Advanced Studies, SOKENDAI, Japan
[2]National Institute of Informatics, Japan
[3]The University of Tokyo, Japan
namgi@nii.ac.jp, {katsuhiko-h, yusuke}@is.s.u-tokyo.ac.jp

Abstract

Many researchers have tried to predict the accuracies of extrinsic evaluation by using intrinsic evaluation to evaluate word embedding. The relationship between intrinsic and extrinsic evaluation, however, has only been studied with simple correlation analysis, which has difficulty capturing complex cause-effect relationships and integrating external factors such as the hyperparameters of word embedding. To tackle this problem, we employ partial least squares path modeling (PLS-PM), a method of structural equation modeling developed for causal analysis. We propose a causal diagram consisting of the evaluation results on the BATS, VecEval, and SentEval datasets, with a causal hypothesis that linguistic knowledge encoded in word embedding contributes to solving downstream tasks. Our PLS-PM models are estimated with 600 word embeddings, and we prove the existence of causal relations between linguistic knowledge evaluated on BATS and the accuracies of downstream tasks evaluated on VecEval and SentEval in our PLS-PM models. Moreover, we show that the PLS-PM models are useful for analyzing the effect of hyperparameters, including the training algorithm, corpus, dimension, and context window, and for validating the effectiveness of intrinsic evaluation.

Keywords: Word Embedding, Intrinsic Evaluation, Extrinsic Evaluation, Structural Equation Modeling, Partial Least Squares Path Modeling

1. Introduction

Word embedding is an indispensable tool for a variety of natural language processing (NLP) tasks, yet it is still unclear why and how it contributes to achieving high accuracy on NLP tasks. A series of extrinsic experiments has proven the effectiveness of word embedding on downstream NLP tasks such as syntactic analysis and semantic textual similarity (Nayak et al., 2016; Conneau and Kiela, 2018). On the other hand, intrinsic evaluation of word embedding, as in word similarity (Bruni et al., 2014; Hill et al., 2015) and word analogy tasks (Mikolov et al., 2013; Gladkova et al., 2016), has been proposed for assessing what linguistic knowledge it encodes. Previous studies (Chiu et al., 2016; Rogers et al., 2018; Wang et al., 2019) tried to prove this intuition by using correlation analysis. Correlation analysis does not, however, assume any causal hypothesis, therefore it is hard to extract any cause-effect relationships from those studies.

Hence, we investigate causal relations between the accuracies of intrinsic and extrinsic evaluation, by applying *partial least squares path modeling* (PLS-PM) (Wold, 1982), a method of structural equation modeling. PLS-PM is a widely accepted method in social science disciplines for analyzing causal relations among observed and latent variables (Henseler et al., 2014). In PLS-PM, hypothetical causal relations are given together via a *causal diagram*, and the strengths of the causal relations are then estimated by fitting the causal diagram to observed data. This method enables analysis of not only the correlations of observed variables but also latent causal relations. Moreover, PLS-PM can incorporate non-metric variables into causal diagrams, such as the training algorithms and corpora.

In this paper, we hypothesize that word embedding encodes some type of linguistic knowledge, such as inflectional morphology and lexicography knowledge, and each type of linguistic knowledge contributes to solving specific categories of downstream tasks. The accuracies of intrinsic and extrinsic evaluation are considered as observed variables, while the types of linguistic knowledge and categories of downstream tasks are regarded as latent variables. We then design causal diagrams to represent hypothetical causal relations among these variables, and we discuss the estimated PLS-PM models using our causal diagrams.

In experiments, we train 600 word embeddings while varying hyperparameters, including training algorithm, corpus, dimension size, and context window. We then measure the accuracies on the BATS dataset (Gladkova et al., 2016) for intrinsic evaluation and on the VecEval (Nayak et al., 2016) and SentEval (Conneau and Kiela, 2018) datasets for extrinsic evaluation. The experimental results reveal several salient relationships between BATS and VecEval/SentEval, the results also demonstrate, however, that the dataset for examining inflectional morphology shows correlation disagreement in the evaluation results. This indicates that the existing intrinsic evaluation for inflectional morphology may not reflect the structure of linguistic knowledge encoded in word embedding. We further investigate the impacts of hyperparameters for training word embedding on intrinsic and extrinsic evaluation, by incorporating hyperparameter categories into causal diagrams.

2. Statistical methodology for testing causal hypotheses

2.1. Background

The relationship between intrinsic and extrinsic evaluation on word embedding has been studied by various researchers (Schnabel et al., 2015; Chiu et al., 2016; Rogers et al., 2018; Wang et al., 2019). They mainly conducted only simple correlation analysis, which only measures the correlation between two sets of observed variables. What NLP researchers want to reveal, however, are causal relationships

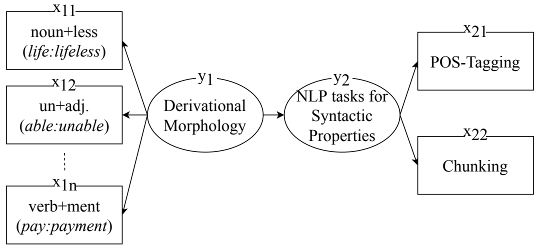

Figure 1: Sample of a causal diagram for the relationship between intrinsic and extrinsic evaluation. A rectangle represents an observed variable, a circle represents a latent variable, and an edge arrow represents a causal relationship.

between intrinsic and extrinsic evaluation, to explain and predict the performance of extrinsic evaluation by intrinsic evaluation. In fact, it is hard to interpret correlation results alone without assuming any causal hypothesis for their existence (Pearl, 2009; Koller and Friedman, 2009). Moreover, with correlation analysis, it is not easy to disentangle the effects of external factors, such as the training algorithm, corpus, and hyperparameters of word embedding, which highly affect the quality of word embedding (Levy et al., 2015; Lai et al., 2016).

To examine causal relations among various observed and latent variables involving analysis of word embedding, we thus require a more general framework for statistical analysis than what correlation analysis provides.

2.2. Structural equation modeling

Structural equation modeling, which was first invented by Wright (1921), provides a convenient framework for statistical analysis that includes several traditional multivariate procedures, such as factor analysis, regression analysis, and canonical correlation analysis. In the social science field, structural equation modeling is used to model and analyze complex relationships between observed and latent variables. In this paper, we adopt structural equation modeling to test a causal hypothesis between intrinsic and extrinsic evaluation on word embedding.

Structural equation modeling estimates a system of linear equations to test the fit of a hypothesized causal model, and thus, its first step involves creating a causal diagram based on prior knowledge. In causal diagrams, rectangles/circles typically represent observed/latent variables, respectively, and edge arrows represent causal relationships between variables.

Figure 1 shows a simple example of a causal diagram. Each observed variable on the left side represents the accuracy of a word embedding for an intrinsic evaluation task on derivational morphology. Those on the right side depict the accuracies for an extrinsic evaluation task on shallow parsing. Two latent variables are designed according to the following hypothesis: The ability to recognize syntactic structures affects the performance of shallow parsing, and this ability largely derives from knowledge on derivational morphology.

Structural equation modeling is separated into two submodels: (1) the *measurement model* has relationships between the observed and latent variables, while (2) the *structural model* consists of the relationships between latent variables. Any causal relationship can be expressed by a linear regression equation, also called a *structural equation*. The measurement model for the diagram in Figure 1 thus consists of the following equations:

$$\begin{array}{ll} x_{11} = \lambda_{11} y_1 + \varepsilon_{11} & x_{21} = \lambda_{21} y_2 + \varepsilon_{21} \\ x_{12} = \lambda_{12} y_1 + \varepsilon_{12} & x_{22} = \lambda_{22} y_2 + \varepsilon_{22} \\ \vdots & \\ x_{1n} = \lambda_{1n} y_1 + \varepsilon_{1n} & \end{array} \tag{1}$$

where the x are observed variables, the y are latent variables, the λ denote weights for each factor, and the ε represent error terms. The structural model also has the following linear equation:

$$y_2 = \beta_{11} y_1 + \zeta_1 \tag{2}$$

where β is a weight and ζ is an error term. Given a causal diagram and the values of observed variables as input, we need to fit such multiple regression equations and latent variables to the input data. After fitting the model, we can interpret the strength of a causal relation from the path coefficient and decide appropriately whether to accept a tested hypothesis according to how well it fits the data.

Various techniques have been developed to estimate model parameters in structural equation modeling. Usually, fitting functions for maximum likelihood estimation are used to fit the system of equations to the variance-covariance matrix of the observed variables, though this method requires that the data be normally distributed and the observations be independent (Jöreskog, 1970). Sometimes those assumptions are unrealistic, however, especially in a case like ours for the accuracies of downstream tasks, which do not usually follow normal distributions. Therefore, in the next section, we introduce a different approach for fitting in structural equation modeling.

2.3. Partial least squares path modeling

Another approach for fitting the model in structural equation modeling is partial least square path modeling (PLS-PM), proposed by Wold (1982). It is often called a component-based approach because it estimates the scores of latent variables from linear combinations of observed variables. PLS-PM does not require difficult assumptions for observed variables, such as a normal distribution and independence (Tenenhaus et al., 2005). Because of the relaxed requirements for observed variables, PLS-PM has been accepted in various social science disciplines as a useful tool for exploratory research (Henseler et al., 2014).

Here, we explain the details of the algorithm for the PLS-PM estimation procedure, following Tenenhaus et al. (2005) and Sanchez (2013). To estimate parameters, PLS-PM first aims to estimate the scores of latent variables. The scores of the latent variables in Figure 1 are thus written as below.

$$y_j = \sum_k w_{jk} x_{jk} + \varepsilon_{jk} \tag{3}$$

Note that PLS-PM does not use or estimate any λ or β before the estimation of y finishes. Because the x are already given as observed variables, we need to estimate the parameter w. PLS-PM thus conducts an iterative procedure for updating w. First, it initializes all w to an arbitrary number that allows the calculated scores of the latent variables to have unit variance. For example, if all w are initialized as 1, then all latent variables in Equation 3 can be estimated as sums of observed variables, as below.

$$\begin{aligned} y_1 &= \textstyle\sum_k x_{1k} + \varepsilon_{1k} \\ y_2 &= \textstyle\sum_k x_{2k} + \varepsilon_{2k} \end{aligned} \tag{4}$$

In the next step, PLS-PM tries to obtain the weights in the structural model, e.g. β_{11} in Equation 2. Note that we do not use the weights of the measurement model, w, in this step. Rather, β is only estimated from the scores of latent variables, by using correlation coefficients between adjacent latent variables. PLS-PM has various options for how to obtain the weights in the structural model. For example, the centroid scheme option obtains β by the following formula:

$$\beta_{ji} = \left(\begin{array}{cc} sign[cor(y_j, y_i)] & \text{if } y_j, y_i \text{ are adjacent} \\ 0 & \text{otherwise} \end{array} \right. , \tag{5}$$

where $sign[a]$ is the sign direction of a, taking a value of ± 1, and $\mathrm{cor}(a, b)$ is the correlation coefficient between a and b. With the obtained β, PLS-PM estimates other scores for the latent variables, y', as below:

$$y'_j = \sum_{i \leftrightarrow j} \beta_i y_i + \zeta_i \tag{6}$$

where $\leftrightarrow$ means that y_i and y_j are connected in the structural model. With these new scores for the latent variables, y', PLS-PM can update the weights of the measurement model, w. In general, it calculates w as a coefficient of ordinary least squares regression on x and y'. The estimation formula for w depends on which variables are the cause; for example, when Equation 3 is given, w will be estimated as below.

$$w_{jk} = ({y'_j}^{\top} y'_j)^{-1} {y'_j}^{\top} x_{jk} \tag{7}$$

PLS-PM then continues the above procedures until w convergences, usually via $|w_{jk}^{e-1} - w_{jk}^{e} < 10^{-5}|$, where e is an epoch number. When the iterative process is complete, PLS-PM has already finished estimating all weights and the scores of the latent variables. Therefore, it can estimate the path coefficients in the structural model and the loadings in the measurement model, which indicate the prediction strength of each path in the PLS-PM model. Here, *path coefficients* in the PLS-PM model are estimated by ordinary least squares regression.

$$\text{Path coefficient}_{ji} = (y_i^{\top} y_i)^{-1} y_i^{\top} y_j \tag{8}$$

A *loading* is usually calculated as the correlation coefficient between an observed variable and a latent variable. During estimation of the path coefficients and loadings, the weights in the measurement model, λ, and the weights in the structural model, β, are also fitted at once. Therefore, this is the end of the PLS-PM fitting process.

To assess a PLS-PM result, researchers use various reliability indexes. First, the design of the measurement model can be examined with *Cronbach's* α and *Dillon–Goldstein's* ρ for internal consistency. Cronbach's α can be interpreted as an average value of inter-variable correlation coefficients. Dillon–Goldstein's ρ is used to examine the composite reliability of the measurement model. In general, both metrics should be larger than 0.7 for unidimensionality of the proposed measurement model.

The structural model is usually interpreted according to its structural equations. Because these equations are estimated by ordinary least squares regression, we can simply validate each equation with a p value. Also, the determination coefficient R^2 and Goodness-of-Fit (GoF) are usually evaluated to assess the quality of the structural model. This evaluation method is similar to other multiple regression analysis methods. In PLS-PM, a latent variable with $R^2 > 0.6$ is considered highly explained. Moreover, a PLS-PM model is considered strong when it achieves a GoF value over 0.7 (Sanchez, 2013).

3. Experimental setup

3.1. Design for casual diagrams incorporating intrinsic and extrinsic evaluation

In this paper, we examine causal hypotheses between intrinsic and extrinsic evaluation of word embedding with the PLS-PM methodology. We thus aim to fit PLS-PM models, following causal hypotheses of previous studies. Our causal hypothesis is that the accuracies of extrinsic evaluation can be explained by the accuracies of intrinsic evaluation with causal relations, as Chiu et al. (2016) assumed. The causal diagram shown in Figure 2 represents this hypothesis.

Following the previous studies (Chiu et al., 2016; Rogers et al., 2018; Wang et al., 2019), we introduce the structure of datasets for intrinsic and extrinsic evaluation to our causal diagram. For intrinsic evaluation, we employ the BATS dataset (Gladkova et al., 2016). We do not use word similarity datasets, because of the ambiguous definition of similarity and the problem of inter-annotator agreement on the dataset (Batchkarov et al., 2016). The BATS dataset consists of four linguistic categories containing ten subcategories, such as *inflectional morphology, derivational morphology, lexicography knowledge*, and *encyclopedia knowledge*. Table 1 lists more details of the BATS dataset. Following Gladkova et al. (2016), we assume that each linguistic category is one latent variable that reflects the accuracies of its ten subcategories for the measurement model in our causal diagrams. By binding subcategories with one latent variable, we can reduce the number of parameters in the PLS-PM model, which allows us to fit the model with fewer samples. Moreover, we can examine whether the structure of linguistic knowledge on the BATS dataset can be applied to word embedding, by investigating the reliability of the measurement model. Lastly, we use the vector offset method (Mikolov et al., 2013) to solve the BATS dataset, well known as the *Man + King = Woman + ?*. Because we intend to avoid the effectiveness of machine learning methods, such as the LRCos method (Gladkova et al., 2016), for evaluating how well linguistic knowledge is embedded.

For extrinsic evaluation, we employ the VecEval (Nayak et al., 2016) and SentEval (Conneau and Kiela, 2018)

dataset-category (latent variable)	tasks (observed variables)
BATS-Inflectional Morphology (INF)	regular plurals, plurals (orthographic changes), comparative degree, superlative degree, infinitive:3ps.sg, infinitive:participle, infinitive:past, participle:3ps.sg, participle:past, 3ps.sg:past
BATS-Derivational Morphology (DER)	noun+less, un+adj., adj.+ly, over+adj./ved, adj.+ness, re+verb, verb+able, verb+er, verb+ation, verb+ment
BATS-Lexicography Knowledge (LEX)	hypernyms (animals), hypernyms (miscellaneous), hyponyms (miscellaneous), meronyms (substance), meronyms (member), meronyms (part-whole), synonyms (intensity), synonyms (exact), antonyms (gradable), antonyms (binary)
BATS-Encyclopedia Knowledge (ENC)	geography (capitals), geography (country:language), geography (uk city:county), people (nationalities), people (occupation), animals (the young), animals (sounds), animals (shelter), other (thing:color), other (male:female)
VecEval-Syntactic Properties (SYN)	POS-tagging (Toutanova et al., 2003), Chunking (Sang and Buchholz, 2000)
VecEval-Semantic Properties (SEM)	Named Entity Recognition (Sang and Erik, 2002), Sentiment Classification (Socher et al., 2013), Question Classification (Li and Roth, 2006), Natural Language Inference (Ganitkevitch et al., 2013)
SentEval-Classification (CLA)	Movie Review, Product Review, Subjectivity Status, Opinion-polarity (Wang and Manning, 2012), Binary Sentiment Analysis, Fine-grained Sentiment Analysis (Socher et al., 2013), Question Classification (Li and Roth, 2006)
SentEval-Natural Language Inference (NLI)	Natural Language Inference (Marelli et al., 2014)
SentEval-Semantic Textual Similarity (STS)	STS 2012 (Agirre et al., 2012), STS 2013 (Agirre et al., 2013), STS 2014 (Agirre et al., 2014), STS 2015 (Agirre et al., 2015), STS 2016 (Agirre et al., 2016), STS Benchmark (Cer et al., 2017), SICK-R (Marelli et al., 2014)
SentEval-Paraphrase Detection (PD)	Paraphrase Detection (Dolan et al., 2004)

Table 1: Details of the datasets used for our PLS-PM models.

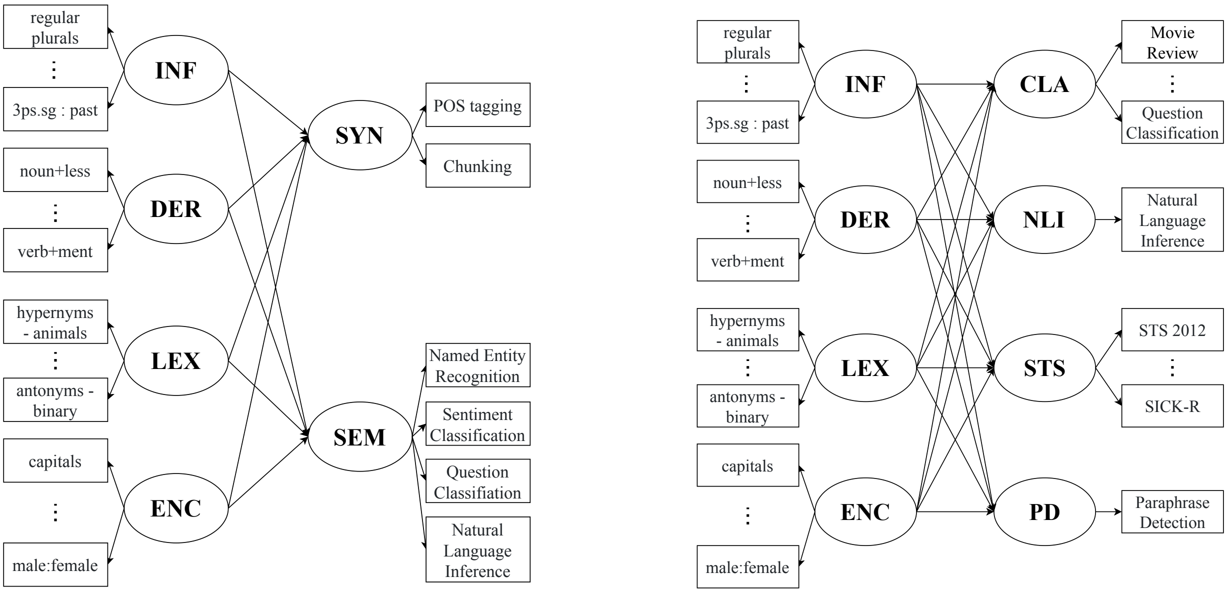

Figure 2: Causal diagrams for BATS-VecEval (left) and BATS-SentEval (right). All abbreviations are defined in Table 1.

datasets. Those authors classified their tasks into NLP research areas, such as tasks for *syntactic* and *semantic properties* in VecEval, and *classification, natural language inference, semantic textual similarity*, and *paraphrase detection* in SentEval, as listed in Table 1. We design latent variables for extrinsic evaluation with the structures of VecEval and SentEval in the same way as for BATS. For example, the latent variable for syntactic properties has the accuracies of POS tagging and chunking as observed variables. Table 1 lists details for the latent and observed variables from BATS, VecEval, and SentEval. Hereafter, we refer to the PLS-PM model using the BATS and VecEval datasets as BATS-VecEval, and to the one using the BATS and SentEval datasets as BATS-SentEval.

Note that the downstream tasks in VecEval and SentEval use various performance indicators, such as the accuracy, F1 score, and Pearson's r. However, we do not unify or transform them, because we need its own performance indicator of each dataset as suggested by the original papers. Therefore, we do not change the values of indicators except through normalization. Furthermore, we distinguish the two causal diagrams for VecEval and SentEval, and do not merge them. The main reason is that they use different neural network models for solving downstream tasks. We should avoid model effects for observed variables, because we do not consider any effect of a machine learning model

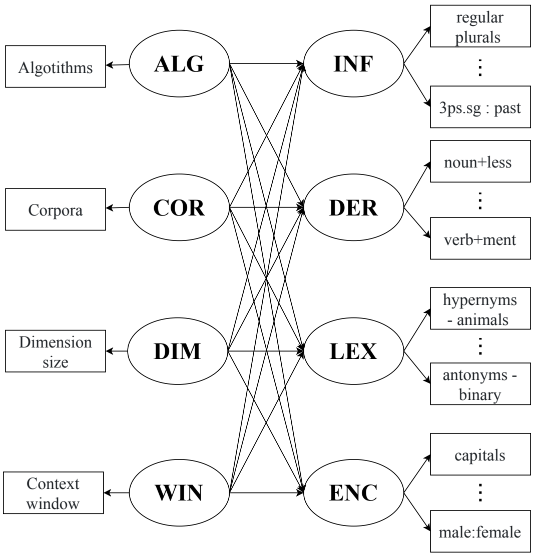

Figure 3: Causal diagram for hyperparam-BATS. All abbreviations are defined in Tables 1 and 2.

in our PLS-PM models.

3.2. Design for causal diagrams incorporation hyperparameters

In addition, we also incorporate hyperparameters in our PLS-PM models. Hyperparameters obviously have a strong effect on the accuracies of intrinsic and extrinsic evaluation (Levy et al., 2015; Lai et al., 2016), although they only conducted correlation analysis. In this paper, we investigate the effectiveness of the hyperparameters of word embedding, including the training algorithm, corpus, dimension, and context window. As a result, we suggest another causal diagram consisting of intrinsic and extrinsic evaluation and hyperparameters for word embedding, as shown in Figure 3 and 4. Note that the hyperparameters are independent variables with respect to each other, and we do not bind them as one latent variable. Moreover, because they include non-metric variables such as the algorithm and corpus, we use transformed scores of the hyperparameters during PLS-PM estimation, following Russolillo (2012). We refer to the PLS-PM model for the above causal diagram as hyperparam-BATS.

Furthermore, we incorporate hyperparameters into BATS-VecEval and BATS-SentEval as illustrated in Figure 4. It is important that we do not directly connect the latent variables of hyperparameters with the latent variables of extrinsic evaluation in our causal diagram. In other words, we assume that the effectiveness of hyperparameters for extrinsic evaluation can be explained only through the accuracies of intrinsic evaluation, which implies the ability of linguistic knowledge. Our causal diagram follows the ideal assumption that intrinsic evaluation namely, that intrinsic evaluation examines the general quality of word embedding; therefore, it should also predict the accuracies of extrinsic evaluation (Chiu et al., 2016). We aim to examine this hypothesis with our PLS-PM models using the above causal diagrams.

algorithm (ALG)	CBOW, Skipgram, Fasttext
corpus (COR)	Wikipedia, New York Times
dimension size (DIM)	50, 100, 150, 200, 250, 300, 350, 400, 450, 500
context window (WIN)	1, 3, 5, 7, 9, 11, 13, 15, 17, 19

Table 2: Hyperparameters for training word embeddings.

	INF	DER	LEX	ENC	R^2
SYN	-	0.773	1.310	-	0.656
SEM	-	-0.189	-	0.771	0.546

Table 3: Path coefficients for each path and R^2 for the endogenous latent variables on BATS-VecEval. Paths with $p > 0.05$ are omitted.

3.3. Sampling word embedding models

To fit a PLS-PM model, not only the causal diagrams but also the values of observed variables are required as input. In general, structural equation modeling demands sufficient sample data for fitting, usually more than 200 samples (Kline, 2015). Therefore, we train a number of word embeddings with various sets of hyperparameters, according to previous studies (Levy et al., 2015; Chiu et al., 2016; Rogers et al., 2018). Table 2 lists the hyperparameters used for increasing the number of word embeddings. As a result, we obtain 600 word embeddings. Because the hyperparameters of word embedding have already been reported to affect the accuracies of intrinsic and extrinsic evaluation significantly (Levy et al., 2015; Lai et al., 2016), we regard the result of one task with one word embedding as one data sample. As a result, our observed variable is a 600-dimension vector consisting of the results of BATS, VecEval, and SentEval on 600 word embeddings.

We use the R package `plspm`[1] for our experiments, and for reproducibility we share our experimental scripts and all observed variable data at `https://github.com/mynlp/embedding-evaluation-plspm`.

4. Results and discussions

4.1. Relationship between intrinsic and extrinsic evaluation

4.1.1. BATS-VecEval

First, we examine the reliability of BATS-VecEval. Cronbach's α and Dillon–Goldstein's ρ, for validating the measurement model of BATS-VecEval, are both larger than 0.7, indicating that the measurement model of BATS-VecEval is acceptable. The *GoF* of BATS-VecEval is 0.6484, which is also considered an acceptable value (Akter et al., 2011). Therefore, we can accept BATS-VecEval and its causal hypothesis.

We can interpret the effectiveness of a path between latent variables with the path coefficient. In BATS-VecEval, there are eight paths between intrinsic and extrinsic evaluation. Table 3 lists their coefficients and the R^2 values for SYN (VecEval-Syntactic Properties) and SEM (VecEval-Semantic Properties). Four paths, namely, DER (BATS-Derivational Morphology)-SYN, DER-SEM, LEX

[1] `https://github.com/gastonstat/plspm`

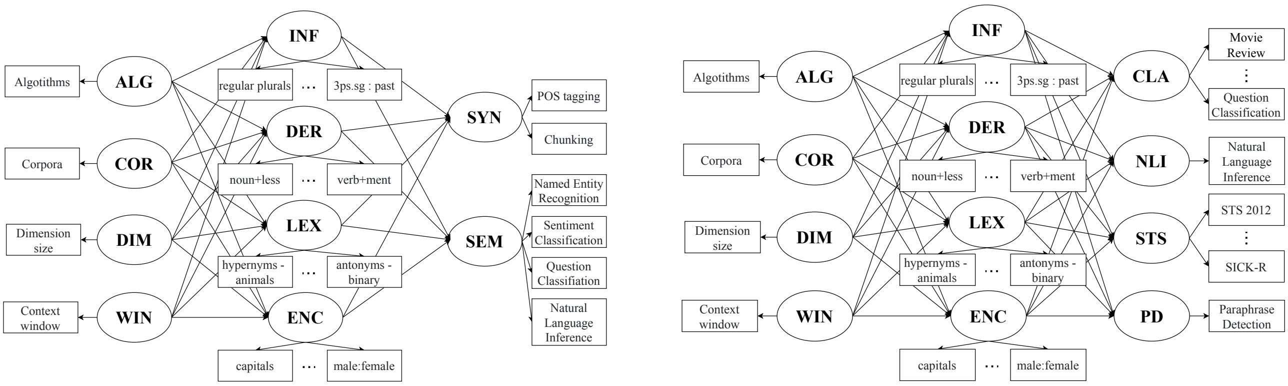

Figure 4: Causal diagrams for hyperparam-BATS-VecEval (left) and hyperparam-BATS-SentEval (right). All abbreviations are defined in Tables 1 and 2.

(BATS-Lexicography Knowledge)-SYN, and ENC (BATS-Encyclopedia Knowledge)-SEM, have $p < 0.05$, indicating significant causal relations. The high path coefficients for DER-SYN and ENC-SEM are intuitively understandable, because knowledge of derivational morphology helps syntactic analysis tasks such as POS tagging, and encyclopedia knowledge is indispensable in semantic analysis. The relation between lexicography and syntax is not trivial, but it has already been reported that accuracy on SimLex-999 (Hill et al., 2015), a dataset of word similarity to distinguish lexicographical relations, is correlated with POS tagging and chunking (Chiu et al., 2016). Our result is consistent with that observation. Another interesting observation is that INF, the latent variable for inflectional morphology, does not have any significant effect on the downstream tasks in VecEval. We further discuss inflectional morphology in Section 4.3. Among the rejected paths, the rejection of the path between lexicography knowledge and tasks of semantic properties seems counter-intuitive. We hypothesize that the main reason derives from the components of SEM; amed entity recognition, sentiment classification, question classification and natural language inference. It is understandable that lexicography knowledge may not have enough explanatory power for some tasks for the SEM latent variable, such as named entity recognition. Also, this can explain why the R^2 value of SYN is higher than that of SEM, because of the similarity of the tasks for the SYN latent variable.

4.1.2. BATS-SentEval

Next, we investigate BATS-SentEval in the same way. For the measurement model, both Cronbach's α and Dillon–Goldstein's ρ are larger than 0.7, indicating that the assumption of the causal diagram between the observed and latent variables is acceptable. The GoF of BATS-SentEval is 0.711, which is higher than that of BATS-SentEval. This implies that the accuracies of BATS can better explain the accuracies of SentEval than those of VecEval. Therefore, we conclude that BATS-SentEval is also acceptable.

As listed in Table 4, in the structural model of BATS-SentEval, all paths are accepted with $p < 0.05$, except INF (BATS-Inflectional Morphology) -NLI (SentEval-Natural Language Inference) and INF-STS (SentEval-Semantic Textual Similarity). The results show that ENC, for encyclopedia knowledge, shows high path coefficients with all latent variables for the SentEval dataset, as SEM shows for VecEval dataset. Among the latent variables of SentEval, classification tasks are well explained with derivational morphology, lexicography knowledge, and encyclopedia knowledge. Because most tasks of the CLA (SentEval-Classification) latent variable consist of sentiment analysis, this may indicate that such linguistic knowledge is useful for sentiment analysis tasks. The results also show, however, that NLI and STS are the best explained latent variables by the accuracies of BATS, according to the R^2 values. When $R^2 > 0.8$, it indicates that an endogenous latent variable is excellently explained by its independent latent variables. Therefore, we argue that encyclopedia knowledge is strong enough to explain the evaluation results of semantic textual similarity, while the path coefficients of DER-STS and LEX-STS are low. In contrast, PD (SentEval-Paraphrase Detection) shows the lowest R^2 value in BATS-SentEval. Although the value is not under the cutoff for rejecting this latent variable, it may indicate that the paraphrase detection task is not well explained by the accuracies of BATS.

	INF	DER	LEX	ENC	R^2
CLA	-0.565	1.140	1.490	0.716	0.619
NLI	-	0.368	0.640	0.647	0.807
STS	-	-0.397	-0.216	0.837	0.874
PD	-0.358	-0.812	-0.321	0.448	0.482

Table 4: Path coefficients for each path and R^2 for the endogenous latent variables on BATS-SentEval. Paths with $p > 0.05$ are omitted.

4.2. Impact of hyperparameters

As Levy et al. (2015) and Lai et al. (2016) reported, hyperparameters for the training of word embedding affect the accuracies on intrinsic and extrinsic evaluation. We thus analyze the effect of hyperparameters by adding new latent variables for hyperparameter values to the causal diagrams, as shown in Figure 3 and 4.

4.2.1. hyperparam-BATS

First, we examine hyperparam-BATS. Note that ALG, COR, DIM, and WIN consist of one observed variable;

	ALG	COR	DIM	WIN	R^2
INF	-0.312	-0.213	0.580	-0.249	0.541
DER	0.969	-0.031	0.136	-0.068	0.963
LEX	-0.937	-0.106	0.150	-0.060	0.915
ENC	-0.861	0.268	0.218	0.072	0.865

Table 5: Path coefficients for each path and R^2 for the endogenous latent variables on hyperparam-BATS.

therefore, we do not need to validate the measurement model of hyperparam-BATS. Other latent variables, such as INF, DER, LEX, and ENC, show higher Cronbach's α and Dillon–Goldstein's ρ values than 0.7, as with BATS-VecEval and BATS-SentEval. Moreover, the *GoF* of hyperparam-BATS is 0.7521, the best value among our PLS-PM models. Therefore, it is obvious for hyperparam-BATS that the hyperparameters of word embedding are strongly effective for the accuracies of intrinsic evaluation. Table 5 lists that the path coefficients and R^2 values for the structural model on hyperparam-BATS. There is no rejected path with $p > 0.05$, which indicates that all the hyperparameters have some impact on the tasks in the BATS dataset. Training algorithms have especially strong relations with all categories of the BATS dataset, as indicated in the table. All hyperparameter values are processed with the nominal scaling (Russolillo, 2012), as we mentioned at Section 3.2. It means that we can not use the sign of path coefficients for interpretation. Therefore, we can conclude that training algorithm is the strongest factor for explaining the accuracies of intrinsic evaluation on hyparparam-BATS, because of the high intensity of its path coefficient. The other hyperparameters are much weaker for predicting latent variables in term of path coefficients than the training algorithms are. For encyclopedia knowledge, the path coefficients of the corpus and dimension are relatively high. This implies that the accuracies of encyclopedia knowledge are more related to the training corpus and dimension than those of other linguistic knowledge. Meanwhile, most latent variables of intrinsic evaluation have salient R^2 values greater than 0.85, with only the R^2 value for inflectional morphology being low, at 0.541. This problem is investigated in Section 4.3.

4.2.2. hyperparam-BATS-VecEval and hyperparam-BATS-SentEval

Next, we investigate hyperparam-BATS-VecEval and hyperparam-BATS-SentEval to incorporate hyperparameters into the analysis of relationships between intrinsic and extrinsic evaluation. The main causal hypothesis of both hyperparam-BATS-VecEval and hyperparam-BATS-SentEval is that the effectiveness of hyperparameters for word embedding on extrinsic evaluation can be explained through the accuracies of intrinsic evaluation. To validate this hypothesis, we focus on the R^2 and *GoF* values of hyperparam-BATS-VecEval and hyperparam-BATS-SentEval. If our causal hypothesis is helpful in explaining the accuracies of extrinsic evaluation, then we should find that the R^2 and *GoF* values of hyperparam-BATS-VecEval and hyperparam-BATS-SentEval are higher than those of BATS-VecEval and BATS-SentEval.

	ALG	COR	DIM	WIN	R^2
INF	-0.687	0.281	0.353	-0.127	0.691
DER	0.974	-	0.119	-0.051	0.966
LEX	-0.941	-0.061	0.153	-0.050	0.916
ENC	-0.878	0.226	0.212	0.062	0.871
	INF	DER	LEX	ENC	R^2
SYN	-0.335	0.953	1.390	0.497	0.688
SEM	-0.447	-	0.183	0.992	0.578

Table 6: Path coefficients for each path and R^2 for the endogenous latent variables on hyperparam-BATS-VecEval. Paths with $p > 0.05$ are omitted.

Tables 6 and 7 list the path coefficients of hyperparam-BATS-VecEval and hyperparam-BATS-SentEval. For both models, the results show that the R^2 values of most latent variables increase. Specifically, both SYN and SEM in hyperparam-BATS-VecEval have better R^2 values than they do in BATS-VecEval. Moreover, Table 8 lists the *GoF* values for all the PLS-PM models. The *GoF* of hyperparam-BATS-VecEval is 0.7445, showing salient improvement over the value for BATS-VecEval, 0.6484. Therefore, we can conclude that extrinsic evaluation in VecEval is more explainable with our causal hypothesis on hyperparam-BATS-VecEval.

On the other hand, it may be difficult to accept the same conclusion as that for hyperparam-BATS-VecEval on hyperparam-BATS-SentEval. As listed in Table 7, the R^2 values of CLA and STS on hyperparam-BATS-SentEval decrease below those on BATS-SentEval. This indicates that the design of hyperparam-BATS-SentEval is not useful for explaining many tasks of extrinsic evaluation in SentEval. Though the *GoF* of hyperparam-BATS-SentEval is higher than that of BATS-SentEval, this result highly depends on the structural equations between the hyperparameters and BATS, not on those between BATS and SentEval. As a result, our causal hypothesis, the effectiveness of hyperparameters for extrinsic evaluation can be explained only through the accuracies of intrinsic evaluation, is not useful for explaining the accuracies of SentEval dataset. This result implies two possible interpretation: that the accuracies of extrinsic evaluation can be explained directly by the hyperparameters, or that the tasks of intrinsic evaluation in BATS are not sufficient to explain the accuracies of the task in the SentEval dataset. We leave this discussion as future work.

4.3. Discussion with respect to previous studies

Our analysis using PLS-PM reveals that the accuracies on intrinsic evaluation, for BATS, can explain the accuracies on extrinsic evaluation, for VecEval and SentEval. Some of these relations were already reported in previous literature using correlation analysis. For example, POS tagging and chunking (VecEval-SYN) are clearly helped by derivational morphology and lexicography knowledge, which was reported in Chiu et al. (2016), Rogers et al. (2018), and Wang et al. (2019). Similarly, classification and natural language inference tasks require derivational morphology, lexicog-

	ALG	COR	DIM	WIN	R^2
INF	-0.456	-	0.540	-0.210	0.545
DER	0.976	-	0.113	-0.043	0.967
LEX	-0.943	-0.059	0.147	-0.045	0.917
ENC	-0.888	0.196	0.207	0.063	0.874
	INF	DER	LEX	ENC	R^2
CLA	-	0.991	1.300	0.204	0.579
NLI	-	0.232	0.516	0.545	0.810
STS	-	-0.455	-0.190	0.689	0.871
PD	-0.555	-	0.430	0.282	0.522

Table 7: Path coefficients for each path and R^2 for the endogenous latent variables on hyperparam-BATS-SentEval. Paths with $p > 0.05$ are omitted.

PLS-PM model	Goodness-of-Fit
BATS-VecEval	0.6484
BATS-SentEval	0.7110
hyperparam-BATS	0.7521
hyperparam-VecEval	0.7445
hyperparam-SentEval	0.7495

Table 8: *GoF* values for our PLS-PM models.

raphy knowledge, and encyclopedia knowledge, which was also reported in Rogers et al. (2018) and Wang et al. (2019). Meanwhile, our PLS-PM models also suggest some counter-intuitive relations between intrinsic and extrinsic evaluation. We already explained the reasons for some results that conflict with those of previous studies, such as the lexicography knowledge and NLP tasks for semantic properties in VecEval. The largest problem is that, in this paper, the latent variable of inflectional morphology shows many rejected structural equations with $p > 0.05$, negative path coefficients on accepted structural equations, and relatively low R^2 values in the overall PLS-PM models. This indicates that the accuracies of inflectional morphology on the BATS dataset may not have sufficient explanatory power for extrinsic evaluation. This result conflicts with the results of previous studies, which reported that the accuracies of inflectional morphology correlate with the accuracies of extrinsic evaluation (Rogers et al., 2018; Wang et al., 2019).

This issue can be explained by the following reasons. First, we suppose that differences in the experimental setting for word embedding lead to conflicting results on inflectional morphology. For example, the accuracies of inflectional morphology in previous studies were calculated by the LR-Cos method (Gladkova et al., 2016), which differs from our experimental setup. In addition, the sample space of word embedding also differs, especially the conditions of the training algorithm and corpus. We leave further analysis on the effectiveness of those differences in our PLS-PM models for future work.

Finally, we also investigate the relationships between the subcategories of inflectional morphology and the estimated score of INF in our PLS-PM models. The left side of Figure 5 shows a plot with the loading of the observed variables, which is a correlation coefficient between the scores of latent and observed variables. The results show that some observed variables have negative loadings, which indicates that subcategories of inflectional morphology on BATS may not correlate well with each other. We can find the same problem in correlation analysis among the observed variables of INF, as shown on the right side of Figure 5. The accuracies for noun plural questions and degrees of adjective inflection obviously do not correlate well. This implies that word embedding may encode the inflectional morphology for nouns and for adjectives in different ways, unlike the structure of the BATS dataset. Therefore, we assume this as the main reason why the INF latent variable is not estimated well in our PLS-PM models.

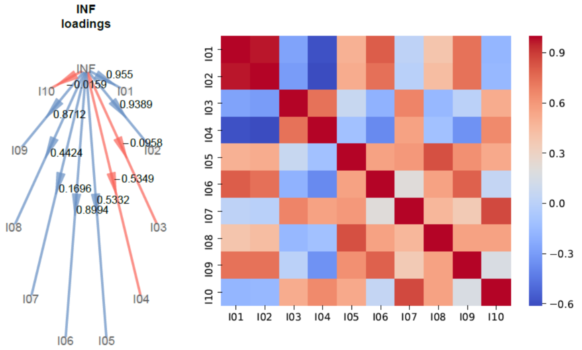

Figure 5: (left) Loading plot of the observed variables for the INF latent variable. A red arrow indicates a negative loading. (right) Spearman correlation heatmap for the INF questions in BATS. Here, I01 and I02 are noun plural questions, I03 and I04 are degrees of adjective inflection, and the other questions are about verbs.

5. Conclusion

In this paper, we employ the PLS-PM method to determine comprehensive relations with causal diagrams composited with intrinsic and extrinsic evaluation, focusing on word embedding. We have found that our PLS-PM models enable statistical analysis that is hard for correlation analysis, such as verifying the existence of causal relations between intrinsic and extrinsic evaluation, the explanatory power of intrinsic evaluation for extrinsic evaluation, and the effectiveness of hyperparameters on intrinsic and extrinsic evaluation. As a result, we have proven part of a causal hypothesis in previous studies, namely, that the accuracies of intrinsic evaluation can explain the accuracies of extrinsic evaluation. In addition, our PLS-PM models have provided novel findings, such as the structural problem of inflection knowledge in the BATS dataset.

Camacho-Collados and Navigli (2016) argued that previous studies on relations between intrinsic and extrinsic evaluation have salient limitations in terms of generality. We believe that our contribution is to employ a statistical methodology to investigate causal relations between intrinsic and extrinsic evaluation, in order to prove them with more generality. In future work, we hope to apply PLS-PM analysis to other vector representations, such as contextualized word representations (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018; Radford et al., 2019) for more general insight about word embedding.

6. Bibliographical References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Akter, S., D'Ambra, J., and Ray, P. (2011). An evaluation of PLS based complex models: the roles of power analysis, predictive relevance and gof index. In *A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, Michigan, USA, August 4-8 2011*.

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12. Association for Computational Linguistics.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Camacho-Collados, J. and Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 43–50, Berlin, Germany.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

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*, pages 1–6.

Conneau, A. and Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

Henseler, J., Dijkstra, T., Sarstedt, M., Ringle, C., Diamantopoulos, A., Straub, D., Jr, D., Hair, J., Hult, T., and Calantone, R. (2014). Common beliefs and reality about pls: Comments on rönkkö & evermann (2013). *Organizational Research Methods*, 17:182–209, 04.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lai, S., Liu, K., He, S., and Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, X. and Roth, D. (2006). Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

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*, pages 746–751.

Nayak, N., Angeli, G., and Manning, C. D. (2016). Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23.

Pearl, J. (2009). *Causality*. Cambridge university press.

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*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

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*, pages 2690–2703, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Russolillo, G. (2012). Non-metric partial least squares. *Electronic Journal of Statistics*, 6:1641–1669.

Sanchez, G. (2013). Pls path modeling with r. *Berkeley: Trowchez Editions*, 383:2013.

Sang, E. F. T. K. and Buchholz, S. (2000). Introduction to the conll-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Sang, T. K. and Erik, F. (2002). Introduction to the conll-2002 shared task: language-independent named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., and Lauro, C. (2005). Pls path modeling. *Computational statistics & data analysis*, 48(1):159–205.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for computational Linguistics.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pages 90–94. Association for Computational Linguistics.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19.

Wold, H. (1982). Soft modeling: the basic design and some extensions. *Systems under indirect observation*, 2:343.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585.