IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 4, 2019, accepted November 21, 2019, date of publication December 12, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957812

Inference of Biomedical Relations Among Chemicals, Genes, Diseases, and Symptoms Using Knowledge Representation Learning

WONJUN CHOI AND HYUNJU LEE
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Hyunju Lee (hyunjulee@gist.ac.kr)

This work was supported in part by the Bio-Synergy Research Project of the Ministry of Science and ICT through the National Research Foundation of Korea (NRF) under Grant NRF-2016M3A9C4939665, in part by the NRF of Korea Grant Funded by the Korean Government (MSIT) under Grant NRF-2018M3A9A7053266, and in part by a grant of the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health&Welfare, South Korea, under Grant HI18C0460.

ABSTRACT Knowledge representation learning represents entities and relations of knowledge graph in a continuous low-dimensional semantic space. Recently, various representation learning models have successfully been developed to infer novel relations in general-purpose knowledge bases such as FreeBase and WordNet. However, few studies have used such models for biomedical data for inferring useful relations among biomedical entities such as genes, chemicals, diseases, and symptoms. This study aimed to compare the potential of representation learning models in extracting biomedical relations by using four different types of representation learning models, viz., TransE, PTransE, TransR, and TransH. For training and evaluating the models, we collected and utilized manually curated data from public databases, including relations among chemicals, genes, diseases, and symptoms. Overall, TransE, the most efficient translation-based monolingual knowledge graph embedding model, displayed the best performance with a higher learning speed for large-scale biomedical data. Using TransE, we inferred new relations among chemicals, genes, diseases, and symptoms, and evaluated the reliability of these inferred relations. Furthermore, TransE outperformed an existing statistical method used in the Comparative Toxicogenomics Database for inferring new chemical-disease relations. Together, the present results show that the representation learning model is useful for inferring new biological data from numerous existing biomedical data.

INDEX TERMS Knowledge representation learning, biomedical knowledge graph.

I. INTRODUCTION

Multi-relational data contained in common knowledge bases (KBs) are often represented using knowledge graphs [1], where nodes indicate entities and edges represent the relations linking the entities. Recently, these entities and relations have been represented as vectors using representation learning models such as TransE [2], PTransE [3], TransR [4], and TransH [5], which are specialized for embedding multi-relational data in a low-dimensional vector space. These representation learning models have been widely used in statistical relational learning and help infer new knowledge in many applications including recommender systems, semantic web, and natural language processing [6]. Recently, the contents and volumes of general-purpose KBs such as FreeBase [7] and WordNet [8] have been rapidly expanding owing to collaborative contributions by experts and the public. Accordingly, several studies have focused on improving existing representation learning models on the basis of these general-purpose KBs. However, few studies have applied representation learning models to infer biomedical relations such as chemical-gene, disease-gene, chemical-disease,

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

gene-gene, and disease-symptom relations, despite the large and growing amount of publicly available biomedical multi-relational databases.

Previous studies have attempted to infer new biological knowledge using public resources from biomedical databases and the literature. For example, HerDing [9] is a herb recommendation system for treating diseases on the basis of resources compiled from public databases and the literature. In HerDing, data regarding chemical-gene relations are obtained from two public databases: the Comparative Toxicogenomics Database (CTD) [10] and TCMID [11]. Data regarding gene-disease relations are obtained from MalaCards [12]. Moreover, Swanson's ABC model [13] was used to infer chemical-disease relations, because genes can serve as links between chemicals and diseases. ChemDis [14] is another integrated chemical-disease inference system based on chemical-gene interactions, which involves a hypergeometric test to combine chemical-gene interactions from STITCH [15] and gene-disease relations from Disease Ontology [16] and Disease Ontology Lite [17] for inferring chemical-disease relations. In the CTD [10], manually curated chemical-gene and gene-disease relations are used to infer chemical-disease relations on the basis of variants obtained from a hypergeometric test.

However, these common gene-based approaches may yield various false-positive findings because the gene activation and regulation differ in accordance with the type of disease considered. We hypothesized that representation learning models can adequately help reduce these false-positive findings, because these models learn vector representations of knowledge graphs by reflecting complex relations among entities, and the plausibility of a certain knowledge within the graph is determined through algebraic operations in the low-dimensional vector space.

To assess the feasibility of this approach, in the present study, we first constructed a large-scale biomedical multi-relational dataset containing information on chemicals, genes, diseases, and symptoms from various public databases such as the CTD [10], MalaCards [12], and BioGrid [18], which were converted into an appropriate data format to be used in a representation learning model. By applying these multi-relational data to the representation learning model, we inferred novel chemical-gene, chemical-disease, disease-gene, gene-gene, and disease-symptom relations and evaluated the reliability of new inferred relations. Thereafter, we assessed the performance of four different types of representation learning models, TransE, PTransE, TransR, and TransH on the basis of the biomedical datasets and further compared our approach with another conventional inference approach.

The principal findings of this study work are summarized as follows:

1) We construct a heterogeneous biomedical knowledge graph using manually curated data from various public databases.
2) We conducted several experiments to prove that the representation learning model is very useful to infer new biomedical relationships and has scope for improvement.

The rest of this article is organized as follows. Section II introduces related databases and existing representation learning models used in this study. Section III describes how novel biomedical relations were inferred using representation learning models and how these models were evaluated. Subsequently, our results are further discussed in Section IV. Section V describes the conclusion and future prospects for research on this topic.

II. RELATED MATERIALS

A. DATABASE

1) PubMed DATABASE
PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) is the most widely used database for searching biomedical literature from MEDLINE and life sciences journals. Currently, PubMed contains over 28 million biomedical abstracts, which contain numerous biomedical entities. We extracted the PubMed abstracts and calculated the number of co-occurrence between biomedical entities.
2) COMPARATIVE TOXICOGENOMICS DATABASE
CTD [10] is a publicly available and an important resource and scientific tool for researchers from all biomedical fields. The database provides manually curated biomedical data and inferred data. We used the curated data including chemical-gene, chemical-disease, and disease-gene relations.
3) THE BIOLOGICAL GENERAL REPOSITORY FOR INTERACTION DATASETS
BioGRID [18] is a well-known database of protein/gene interactions manually curated from Medline literature. We used the database to extract gene-gene interactions.
4) THE MalaCards HUMAN DISEASE DATABASE
MalaCards [12] is a comprehensive disease database containing integrated information regarding 72 meta-resources and over 19,000 disease entries. We used the database to obtain disease-symptom relations.

B. REPRESENTATION LEARNING MODEL

A representation learning model or knowledge graph-embedding model is a prominent method for link prediction. A knowledge graph comprises multi-relational data with entities as nodes and relations as edges. The model then embeds entities of a knowledge graph into a continuous low-dimensional space to be represented as vectors, and further embeds the included relations as vectors. In the representation learning model, data are presented as triplets *(i.e., Head Entity, Relation, Tail Entity)*, where *Relation* indicates a relation between the *Head Entity* and *Tail Entity* (e.g., (Steve Jobs, Founded, Apple company)). Herein, we denote a

triplet as *(h, r, t)* and their corresponding vectors as **h**, **r**, and **t**, respectively. A knowledge graph-embedding model helps predict new relations among entities on the basis of existing triplets. In the present study, we applied four different types of representation learning models: TransE, PTransE, TransR, and TransH.

TransE [2] is an energy-based model for learning low-dimensional embeddings of entities. The basic concept of TransE is that the relation between two entities corresponds to the translation between the embeddings of entities. In other words, the embedding of the tail entity t should be close to the embedding of the head entity h plus a vector depending on the relation r (i.e., $\mathbf{h}+\mathbf{r}\approx\mathbf{t}$ for positive triplets). Thus, the energy score function E of TransE (1) is expressed as follows:

$$E(h, r, t) = ||\mathbf{h}+\mathbf{r}-\mathbf{t}||_{L1/L2}, \tag{1}$$

where $\mathbf{h}$, $\mathbf{t}$, and $\mathbf{r} \in \mathbb{R}^d$ are d-dimensional embeddings of h, t, and r, respectively, and satisfy the norm constraints ($||\mathbf{h}||_2^2 = ||\mathbf{t}||_2^2 = 1$). To learn such embeddings, they minimize a margin based-ranking loss function defined as $\mathcal{L} = \Sigma_{(h,r,t)\in\Delta}\Sigma_{(h',r',t')\in\Delta'} max(0, \gamma + E(h, r, t) - E(h', r', t'))$, where Δ represents a set of positive triplets, Δ' indicates a set of corrupted triplets which are triplets with either head or tail replaced by a randomly selected entity from the set of entities, and γ is the margin separating positive and corrupted triplets. The loss function tries to minimize energy scores for the positive triplets and, on the other hand, favors higher values of the energy score function for the corrupted triplets. TransE is simple and efficient, and is therefore easy to train with very large-scale datasets and contains a relatively small number of parameters: $(O(n_e d + n_r d))$, where n_e and n_r are the number of entities and relations and d is the embeddings dimension. Despite its simplicity, TransE performs as adequately as most expressive models with large multi-relational datasets. However, it was reported that TransE has disadvantages for handling multi-step path relations that frequently appear in the general-purpose KBs [3].

PTransE [3] is an extended model of TransE developed to model relation paths for representation learning of KBs and is also known as path-based TransE. The primary difference between TransE and PTransE is that TransE only assesses direct relations between two entities, whereas PTransE considers not only direct relations but also important multi-step path information. For example, $(entity_1, relation_1, entity_2)$ and $(entity_2, relation_2, entity_3)$ can reveal a new relation $(entity_1, relation_1 \circ relation_2, entity_3)$, where $\circ$ is a function that links $relation_1$ and $relation_2$ into a unified relation path representation. As explained in PTransE [3], multiple relation paths are defined as $P(h, t) = \{p_1, \ldots, p_l\}$ connecting two entities h and t, where relation path $p = (r_1, \ldots, r_l)$ represents $h \xrightarrow{r_1} \ldots \xrightarrow{r_l} t$. The energy score function E of PTransE (2) is expressed as follows:

$$E(h, r, t) = E_1(h, r, t) + \frac{1}{Z}\sum_{p\in P(h,t)} R(p|h, t)E(h, p, t), \tag{2}$$

where the first term $E_1(h, r, t)$ is the same as in Equation (1), and the second term represents a function for modeling indirect relations through multiple step relation paths. $R(p|h, t)$ represents the reliability of the relation path p given the entity pair (h, t) and $Z = \Sigma_{p\in P(h,t)}R(p|h, t)$ is a normalization factor, and lastly $E(h, p, t)$ is the energy function of the triplet (h, p, t) concerning the relation path representation.

In contrast with most representation learning models that assume embeddings of entities and relations within the same d-dimensional vector space $\mathbb{R}^d$, TransR [4] assumes that entities and relations are completely different objects; thus, they should not be represented in a common semantic space. Therefore, TransR was proposed as a novel method for modeling entities and relations in different vector spaces such as the entity space and relation space. The energy score function E of TransR (3) is expressed as follows:

$$E(h, r, t) = ||\mathbf{h}M_r + \boldsymbol{r} - \boldsymbol{t}M_r||_{L1/L2}, \tag{3}$$

where M_r is a relation-specific projection matrix that projects entities from the entity vector space to the relation vector space. For a head and tail pair, the relation-specific projection forms the members of a pair actually holding the relation closer with each other, but places the members far away from each other if the pair does not hold the relation. However, TransR tends to not scale well because of its expensive matrix vector operations on very large-size, complex datasets.

Considering the poor performance of TransE [2] for inferring relations with mapping properties such as reflexive, one-to-many, many-to-one, and many-to-many relations among entities despite its efficiency, TransH [5] was developed to overcome these limitations and provide a good trade-off between model capacity and efficiency. TransH enables an entity to have distributed representations for different relations, which indicates different roles of the entity in different relations. Thus, for a particular relation r, the relation-specific translation vector $\mathbf{d}_r$ is positioned in the relation-specific hyperplane $\mathbf{w}_r$ instead of within the same space of entity embeddings. Thereafter, the embedding $\mathbf{h}$ and $\mathbf{t}$ are projected to the hyperplane $\mathbf{w}_r$, and each projection is denoted as follows: $\mathbf{h}_\perp$, $\mathbf{t}_\perp$. In TransH, both projections are expected to be linked via a translation vector $\mathbf{d}_r$ on the hyperplane. Therefore, the energy score function E of TransH (4) can be represented as follows:

$$E(h, r, t) = ||\mathbf{h}_\perp + \mathbf{d}_r + \mathbf{t}_\perp||_{L1/L2}, \tag{4}$$

where $\mathbf{h}_\perp = \mathbf{h}$ - $\mathbf{w}_r{}^\intercal\mathbf{h}\mathbf{w}_r$ and $\mathbf{t}_\perp = \mathbf{t}$ - $\mathbf{w}_r{}^\intercal\mathbf{t}\mathbf{w}_r$. Consequently, TransH yields better results than TransE, especially for complex multi-relational data such as FreeBase.

III. METHODOLOGY

In this section, we first show how we constructed our biomedical datasets to use them in the representation learning models. Second, we explain how we inferred new biomedical relations using the models and also evaluated the performance of the models. Lastly, we introduce several experiments

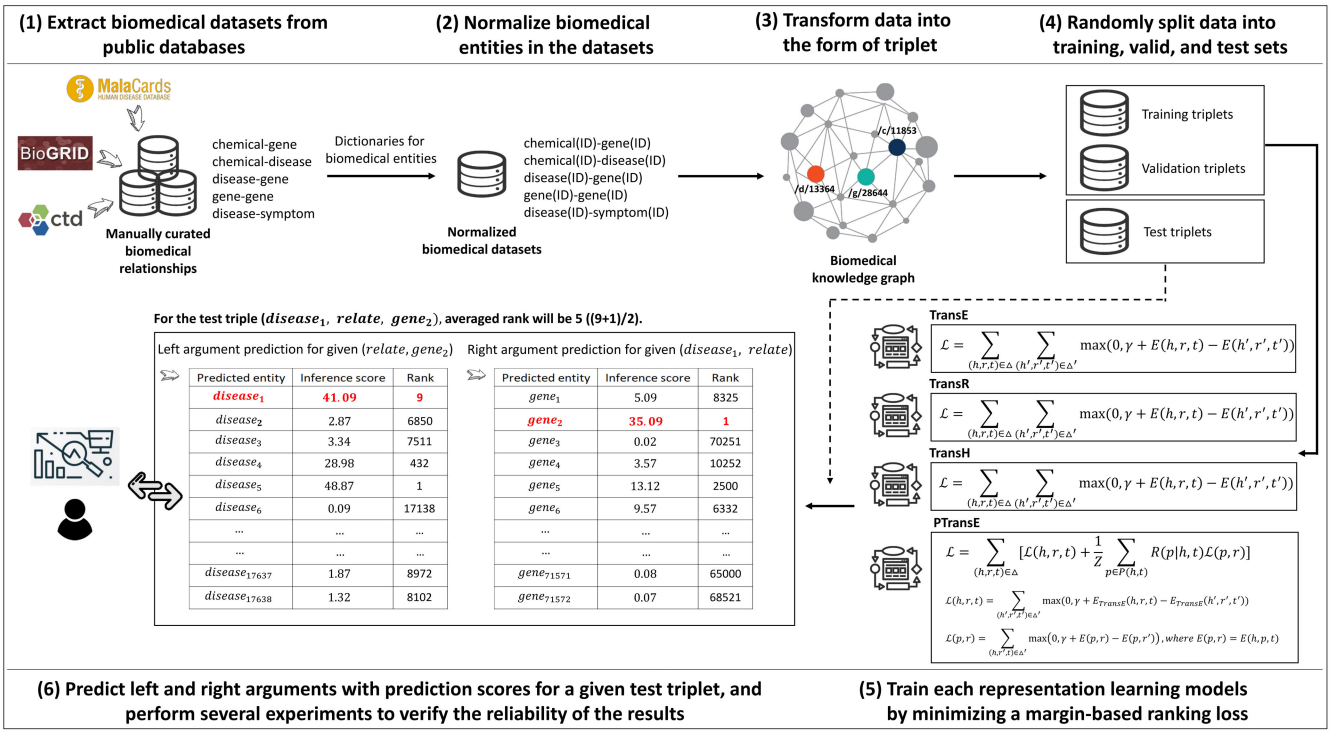

FIGURE 1. An illustration of the entire process of our work. (1) We extracted relations among biomedical entities from public databases; (2) We normalized biomedical entities in the datasets because different databases can use different names for the same entity; (3) We transformed biomedical data into the form of triplet to use them in the representation learning models; (4)–(5) We trained each model by minimizing a loss function using randomly selected training, development and test triplets; and (6) We performed several experiments to verify the reliability of the results from the models.

to verify the reliability of inferred relations using TransE. An illustration of the entire process is shown in Fig 1.

A. CONSTRUCTION OF BIOMEDICAL DATASETS

To infer new relations via the aforementioned training models, we obtained biomedical data, including chemical-gene, chemical-disease, gene-gene, disease-gene, and disease-symptom relations. Relationships were extracted from the CTD, MalaCards, and BioGrid databases, as shown in Table 1, in which the first column represents the relation type between head and tail entities. For example, (chemical, relate, gene) indicates that a chemical upregulates or downregulates a gene, (chemical, relate, disease) represents a chemical that is used to treat or cause a disease, and (disease, relate, gene) indicates a gene targeted for the treatment of a disease or that the gene causes a disease. These three relation types were extracted from the CTD. Moreover, (gene, relate, gene) extracted from BioGrid indicates interactions between two genes. Finally, (disease, have, symptom) relations were extracted from MalaCards. Herein, we only used curated relations owing to their higher confidence levels than inferred relations. The CTD, one of the largest databases indicating relations among chemicals, genes, and diseases, delineates not only biomedical relations that have been manually curated by experts but also includes inferred relations with corresponding inference scores [19].

TABLE 1. Statistics of biomedical knowledge bases obtained from public databases.

Relation types (h, r, t)	Head entities	Tail entities	Interactions	Data sources
Chemical, relate, gene	12,439	35,115	834,214	CTD
Chemical, relate, disease	9348	2973	89,457	CTD
Disease, relate, gene	5111	6760	27,363	CTD
Gene, relate, gene	49,590	49,590	2,193,026	BioGrid
Disease, have, symptom	9060	8728	129,155	MalaCards

Different public databases use different names for the same diseases. For example, in the CTD, ovarian cancer is named "ovarian neoplasms," whereas in the MalaCards database, it is called "ovarian cancer, somatic." Because such name variations can result in different vectors for the same biological concepts, we normalized the names of diseases in our dataset by first obtaining information about the names of diseases from disease dictionaries with various types of disease identifiers such as MeSH [20], OMIM [21], and ICD [22] from public databases. Thereafter, we mapped the names of all diseases to MeSH, OMIM, and ICD. In the absence of appropriate identifiers for the names of specific diseases, new identifiers were assigned. Furthermore, we normalized gene names, because information regarding genes was obtained from two different public databases (CTD and BioGrid). We first constructed a gene dictionary by compiling synonyms for each gene symbol from the CTD and

BioGrid databases, comprising gene symbols with corresponding identifiers and synonyms. Using this gene dictionary, we denoted all gene names in our biomedical datasets as gene identifiers. Furthermore, we assigned new identifiers for some gene names that were not included in the gene dictionary. Consequently, we compiled 3,273,215 relations and 103,625 entities (chemicals, genes, diseases, and symptoms) among these relations.

We examined the properties of the biomedical data by comparing them with two other KBs: WordNet and Freebase. WordNet is a lexical KB of the English language and contains 40,743 entities, 18 relation types, and 151,442 triplet data. FreeBase is a huge KB of general facts, comprising 14,951 entities, 1345 relation types, and 592,213 triplet data. Both KBs are considered standard datasets used to assess various representation learning models. In [23], it was reported that ''averaged triplet number per entity (ATPE)'' is a measurement of diversity and complexity of datasets. The ATPE is calculated as the number of total triplet data divided by the number of total entities. Thus, more triplets lead to more complex structures of the knowledge graph. Usually, the performance of embedding methods is lower in datasets with a higher ATPE, and the ATPE values for WordNet and FreeBase are 3.71 and 39.61, respectively. The ATPE value for our biomedical datasets with 103,625 entities and 3,273,215 triplets is 31.59, which is similar to that for FreeBase. However, there is a difference between FreeBase/WordNet and the biomedical datasets. First, FreeBase and WordNet contain many transitive relations ($(x, y) \in R$ and $(y, z) \in R$ imply $(x, z) \in R$), whereas the transitivity property does not hold for many entities and relations in the biomedical datasets. Second, FreeBase and WordNet contain relatively numerous and distinct relation types, whereas our biomedical datasets have only five relation types that are semantically related with each other. Thus, we need to investigate whether the performances of the representation learning models are affected by these differences in data sets.

B. INFERRING NEW BIOMEDICAL RELATIONS USING REPRESENTATION LEARNING MODELS

We transformed all biomedical data in Table 1 into the form of triplet *(Head Entity, Relation, Tail Entity)* to render the dataset suitable for use in representation learning models. Head and tail entities are represented as unique identifiers with a specific relation type. For example, the disease-gene relation for *(cardiomyopathies is related to CYCS)* is denoted as the triplet *(/d/13364, d_relate_g, /g/28644)*. Furthermore, the chemical-gene relation *(bisphenol a is related to MMS22L)* is represented as the triplet *(/c/11853, c_relate_g, /g/06856)*. Although both *d_relate_g* and *c_relate_g* relation identifiers have the same relation name, i.e., ''relate,'' in Table 1, they are assigned different relation identifiers because they represent relations among different entity types. Thus, each relation has different embeddings.

After transformation of the biomedical data, we randomly split all the triplets into training, validation, and test datasets, as described in Table 2. To train the representation learning models (TransE, PTransE, TransR, and TransH), we used the training and validation datasets in Table 2. The representation learning model predicts head entities for a particular relation type along with a tail entity and predicts tail entities for a particular relation types along with a head entity. To infer new biomedical relations, we input a pair of elements of triplet (head entity, relation type) or (relation type, tail entity) into the trained models. Thereafter, we predicted tail entities or head entities, respectively, in accordance with the input. For example, if we input (/d/13364, d_relate_g) data into the model, the model outputs tail entities with prediction scores from its energy score function. Note that the prediction scores (or inference scores) in this study were calculated by the product of energy scores and -1 to make the results easy to interpret. Thus, higher prediction scores represent greater accuracy.

TABLE 2. Statistics of training, validation and testing data used for the training representation learning models and for comparing the performance of the models each other based on mean rank scores and Hits@10.

Relation types (*h, r, t*)	Total triplets	Train triplets	Valid triplets	Test triplets
Chemical, relate, gene	834,214	833,464	500	250
Chemical, relate, disease	89,457	88,707	500	250
Disease, relate, gene	27,363	26,613	500	250
Gene, relate, gene	2,193,026	2,192,276	500	250
Disease, have, symptom	129,155	128,405	500	250
Total	3,273,215	3,269,465	2500	1250

C. EVALUATION OF MODEL PERFORMANCE AND VERIFICATION OF INFERRED RELATIONS

KBs such as WordNet and Freebase do not contain negative triplets. Because the knowledge in KBs is incomplete, the triplets that are not in the KBs may not be true negative triplets. Thus, to evaluate the representation learning model using these KBs, typical measurements including the area under the curve between true-positive and false-positive data are unsuitable. Instead, Bordes *et al.* [24] suggested the following evaluation method. As mentioned above, the representation learning model predicts head entities for a particular relation type along with a tail entity, and predicts tail entities for a particular relation type along with a head entity. For each test triplet, the original head entity is replaced by all entities of the same types of entities in KBs in turn, which are referred to as corrupted triplets. Furthermore, the energies of these corrupted triplets and the original triplet are computed by an energy score function and sorted in descending order, which is used to determine the rank of the original triplet. The same procedure is then performed for the original tail entity. The average of the ranks for all test triplets is used as the performance metric of the model. For example, considering the test triplet *(cardiomyopathies, relate, CYCS)*, the left entity, i.e., *cardiomyopathies*, is replaced with all other entities in turn, and the energies for the corrupted triplets are computed. The energy score of the original entity was then

compared with those of the corrupted triplets, and the rank was computed. This procedure would also be completed for the right-hand argument, *CYCS*. Note that the other triplets included in the training and validation sets were not used when calculating the rank.

We measured hits@10 representing the proportion of correct entities ranked in the top 10. For example, to measure hits@10 for head, the e_{head} was removed for each test triplet (e_{head}, r, e_{tail}) and replaced by each of the entities of the test set. Thereafter, we assessed whether the e_{head} is ranked in the top 10 among predictions when prediction scores are sorted in descending order. This procedure was repeated for measuring hits@10 for the tail. We used these two evaluation procedures to compare the performance of representation learning models (TransE, PTransE, TransR, and TransH).

Because we inferred new biomedical relations from representation learning models, it was considered as the primary focus of this study to verify the reliability of inferred relations. Thus, we conducted not only performance evaluation of the models but also several additional experiments. First, we investigated whether the highly ranked entities could be considered more reliable because a higher score by the model represents greater accuracy. We trained and tested the model with a randomly selected subset of the data, and assessed whether the highly ranked triplets were more likely to be included in the remaining data. Second, the CTD provides inference scores for the inferred chemical-disease relations to indicate their reliability. Thus, we applied a statistical model used in the CTD to our biomedical data. Thereafter, we compared the reliability of new chemical-disease relations inferred by the representation learning model with that of those inferred by the statistical model used in the CTD. Finally, the co-occurrence represents the simultaneous occurrence of two entities in the same document, sentence, or phrase, which is a measure of the relative closeness between two entities. Thus, we investigated whether two entities in the highly ranked relation co-occurred in the larger number of abstracts by counting the number of the abstract-level co-occurrences of two entities in each predicted relation. Hence, we first compiled approximately 28 million PubMed abstracts for detecting 103,625 entities listed in Table 2 using LinPipe [25]. Therefore, we used our dictionary containing chemical, gene, disease, and symptom entity names and their synonyms. Consequently, we obtained NER results for the 103,625 entities, which were used to determine the number of co-occurrence PMIDs for the inferred relations.

IV. EXPERIMENTS

A. PERFORMANCE OF VARIOUS REPRESENTATION LEARNING MODELS BASED ON OUR BIOMEDICAL DATA SETS

We initially assessed the performance of four different types of representation learning models (TransE, PTransE, TransR, and TransH) based on our biomedical datasets. For training and evaluating the models, we randomly split the biomedical data shown in Table 1 into train, validation, and test datasets, as shown in Table 2. Each of the four representation learning models (TransE, PTransE, TransR, and TransH) was trained with 3,269,465 training triplets and 2500 validation triplets. For TransE, we followed the default configuration used in the source code as follows: learning rate $\lambda = 0.01$, margin $\gamma = 2$, latent dimension $k = 20$, dissimilarity measure $d = \mathrm{L}_1$ distance, and 500 learning epochs. Furthermore, we followed the default configuration in the source code for the other models (PTransE, TransR, and TransH) as follows: learning rate $\lambda = 0.001$, margin $\gamma = 1$, latent dimension $k = 100$, dissimilarity measure $d = \mathrm{L}_1$ distance, and 1000 learning epochs. For the test data, the scores of tail entities were predicted with given head entities and relation types. Inversely, the scores of head entities were predicted with particular tail entities and relation types. Finally, we ranked the predicted relations in accordance with the scores for each given head or tail entity with a particular relation type.

Table 3 summarizes the comparative performance of TransE, PTransE, TransR, and TransH on the basis of mean rank scores using the datasets shown in Table 2. The first column represents the representation learning models. In the second column, separate data types indicating the performance are shown for each data type. The mean rank for the relation type denoted as ''ALL'' represents the average of mean ranks for the five relation types. The third and fourth columns indicate each mean rank for the head and tail entities. For example, in TransE, the mean rank for head entities in the chemical-relate-gene relation type was 445.1, indicating that the chemicals in the test triplets ranked in the top 2.9% on average out of the total number of chemicals (15,267). The averaged mean ranks for both head and tail entities are also shown. For example, TransE achieved a mean rank of 1642.06 out of 103,625 entities, indicating that when a relation type and one of the arguments in some test triplets are entered in TransE, the model predicts its corresponding left- or right-hand argument in the top 1642.06 rank position on average.

Table 4 summarizes the comparison of the performances of each model on the basis of Hits@10 representing the proportion of correct entities ranked in the top 10. The second and third columns mean each hits@10 for the head and tail entities. For example, in TransE, the hits@10 for head entities was 20.08%, indicating that for 20.08% of 1250 test triplets, head entities were correctly ranked in the top 10. The averaged proportion of both head and tail entities are also shown in the last column.

PTransE, TransR, and TransH were originally developed to overcome the limitations of TransE. However, with our biomedical datasets, TransE outperformed the other representation learning models, as shown in Table 3 and Table 4. Thus, we speculated why TransE works better on the biomedical datasets than PTransE, TransR, and TransH despite TransE having a disadvantage for handling complex relation types. First, PTransE is an extending model of TransE to model a path-based representation. The model incorporated

TABLE 3. Comparison of the performance of different representation learning models based on mean rank scores.

Model	Relation Type	Mean Rank for head (the top %)	Mean Rank for tail (the top %)	Average Mean Rank
TransE	ALL	1448.44	1835.67	1642.06
	chemical, relate, gene	445.1 (2.9%)	3106.42 (4.5%)	1775.76
	chemical, relate, disease	1680.95 (11%)	317.52 (2.8%)	999.24
	gene, relate, gene	2256.77 (3.3%)	2302.09 (3.4%)	2279.43
	disease, relate, gene	1445.03 (12.8%)	2772.06 (4.1%)	2108.54
	disease, have, symptom	1414.34 (12.6%)	680.28 (7.8%)	1047.31
PTransE(add)	ALL	3122.24	4508.8	3815.52
	chemical, relate, gene	570.1 (3.7%)	4684.3 (6.9%)	2627.23
	chemical, relate, disease	1925.6 (12.6%)	585.5 (5.2%)	1255.55
	gene, relate, gene	9862.1 (14.4%)	8719.8 (12.8%)	9290.96
	disease, relate, gene	1553.8 (13.8%)	7515.3 (11%)	4534.52
	disease, has, symptom	1699.6 (15.1%)	1039.1 (11.9%)	1369.34
PTransE(mul)	ALL	5060.76	6468.68	5764.72
	chemical, relate, gene	613.74 (4%)	4744.89 (6.9%)	2679.31
	chemical, relate, disease	2044.62 (13.4%)	606.22 (5.4%)	1325.42
	gene, relate, gene	19,428.98 (28.4%)	17,359.92 (25.4%)	18,394.45
	disease, relate, gene	1491.51 (13.2%)	8488.21 (12.4%)	4989.86
	disease, has, symptom	1724.98 (15.3%)	1144.17 (13.1%)	1434.58
PTransE(rnn)	ALL	7283.4	8626.26	7954.83
	chemical, relate, gene	8783.64 (57.5%)	11015.41 (16.1%)	9899.53
	chemical, relate, disease	3555.65 (23.2%)	3148.04 (27.9%)	3351.84
	gene, relate, gene	17898.39 (26.2%)	17020.07 (24.9%)	17459.23
	disease, relate, gene	4596.63 (40.8%)	11021.9 (16.1%)	7809.26
	disease, has, symptom	1582.67 (14%)	925.87 (10.6%)	1254.27
TransR	ALL	2506.81	9460.27	5983.54
	chemical, relate, gene	2418.8 (15.8%)	37,964 (55.5%)	20,191.41
	chemical, relate, disease	2642.2 (17.3%)	464.2 (4.1%)	1553.2
	gene, relate, gene	4151.3 (6.07%)	3747.3 (5.5%)	3949.3
	disease, relate, gene	1624.7 (14.4%)	4057.6 (5.93%)	2841.15
	disease, has, symptom	1697 (15.1%)	1068.2 (12.2%)	1382.6
TransH	ALL	11,611.43	15,261.13	13,436.28
	chemical, relate, gene	7063.06 (46.3%)	15,289.32 (22.4%)	11,176.19
	chemical, relate, disease	7089.58 (46.4%)	5233.12 (46.5%)	6161.35
	gene, relate, gene	32,825.9 (48%)	31,220.01 (45.7%)	32,022.96
	disease, relate, gene	4899.62 (43.5%)	18,136.3 (26.5%)	11,517.96
	disease, has, symptom	6179 (54.8%)	6426.89 (73.6%)	6302.95

TABLE 4. Comparison of the performance of different representation learning models on the basis of Hits@10 (in %).

Model	Hits@10 for head	Hits@10 for tail	Average Hits@10
TransE	20.08	14.48	17.28
PTransE(add)	12.48	9.92	11.2
PTransE(mul)	13.04	8	10.52
PTransE(rnn)	8.24	8.56	8.4
TransR	9.28	7.92	8.6
TransH	1.52	2.16	1.84

connected relational facts between entity pairs instead of only considering the direct relation between two entities. Thus, PTransE works better with data containing transitive relations. As we previously described, FreeBase and WordNet have many transitive relations. For example, FreeBase contains the following relation types: '*people_born_here*', '*capital*', and '*nationality*' and suppose that we have two triplets: T_1 = (*Samuel Leroy Jackson*, *people_born_here*, *Washington D.C.*) and T_2 = (*Washington D.C.*, *capital*, *the USA*). From these two triplets the transitive relation, T_3 = (*Samuel Leroy Jackson*, *nationality*, *the USA*), can be derived. In a similar perspective, in WordNet, the hyponymy relation has transitivity properties. For example, if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture. On the other hand, transitivity is not always satisfied in biomedical datasets owing to the characteristics of biomedical entities and relation types. Suppose that the BioGrid database contains the following gene interactions: ($gene_1$, *relate*, $gene_2$) and ($gene_2$, *relate*, $gene_3$). However, the relation ($gene_1$, *relate*, $gene_3$) does not always exist in BioGrid.

Second, both TransR and TransH were designed to differently project entities depending on each relation type, meaning that they assign an entity with different representations when involved in various relation types. These two methods outperformed TransE on knowledge bases containing numerous and distinct relation types. As mentioned earlier, FreeBase and WordNet have 1345 and 18 distinct relation types, respectively. In addition, except for some relations holding transitive properties, relations are distinct from each other, having been obtained from several different domains such as business, music, and medicine. However, the number of relation types in our biomedical datasets is only five, and chemical, gene, disease, and symptom are related with each other so that relations among them are also related with each other. Therefore, we used TransE to infer new relations between biomedical entities and to proceed with further experiments, because the training duration of TransE was much shorter than that of other models and it displayed the best performance especially with a large amount of biomedical data.

TABLE 5. Statistics of training and validation data used for evaluating the accuracy of relations inferred by TransE.

Relation types (*h, r, t*)	Train triplets	Valid triplets
Chemical, relate, gene	166,693	500
Chemical, relate, disease	17,741	500
Disease, relate, gene	5473	500
Gene, relate, gene	438,455	500
Disease, have, symptom	25,831	250
Total	654,193	2500

B. EVALUATION OF NEW RELATIONS INFERRED USING TRANSE

The aforementioned evaluation was based on the ranking of predicted entities. During subsequent evaluation, we further investigated whether the highly ranked entities were indeed more reliable for inferring relations. Thus, we trained and tested the model with a randomly selected subset of the data and investigated whether the highly ranked predictions were more likely to be included in the remaining data. For each relation type, we randomly selected 20% of the total triplets and 500 triplets, which were used as training and validation datasets, respectively. Of 3,273,215 triplets, 656,693 (= 654,193 + 2500) triplets were used for training the model. Details of these datasets are provided in Table 5. Thereafter, we selected 250 head entities for each relation type for testing, and predicted their corresponding tail entities. To select these 250 entities, we initially sorted all entities from the train triplets of Table 5 in descending order of the number of occurrences. Thereafter, we selected the top 250 entities for each relation type. Finally, we investigated whether the highly ranked predicted triplets were more likely to occur in the 2,616,522 (= 3,273,215 − 656,693) remaining triplets.

Fig 2 shows the distribution trends of direct matches obtained for each rank range, where a direct match indicates that a predicted relation occurs in the set of the remaining triplets. In Fig 2, x-axis is a rank range between top X and top Y rank, and y-axis is the percentage of the number of direct matches in each rank range. The model predicted up to 3000 tail entities for each test triplet. Thus, if 250 chemicals are entered in the model, it yields 750,000 possible relations. In Fig 2, among the 100 predictions ranked in the range between the top 1 and 100, 11.66% of the predicted chemical-disease relations occurred in the 71,216 (= 89,457 - 18,241) remaining triplets on average and tended to decrease with a reduction in the rank. Furthermore, the rest of the bars for each relation type displayed a similar tendency. Thus, This graph shows that the highly predicted relations determined using TransE were more likely to be included in the remaining triplets for disease-gene, chemical-disease, chemical-gene, gene-gene, and disease-symptom relations.

C. RELIABILITY OF THE NEW CHEMICAL-DISEASE RELATIONS INFERRED USING TRANSE COMPARED TO THAT OF THOSE INFERRED USING THE STATISTICAL MODEL USED IN CTD

The CTD provides manually curated chemical-gene, gene-disease, and chemical-disease relations. Further, they provide inferred relations between chemicals and diseases, which are generated by combining chemical-gene data with gene-disease data by using common genes linking chemical and disease entities. The CTD also provides inference scores for the inferred chemical-disease relations to indicate the reliability of the inferred data, which were analyzed using five statistical metrics including the hypergeometric clustering coefficient (C_{XY}), two common neighbor statistics (P_1, P_2), and two novel variants of these metrics (S_{XYA}, W_{XYA}). Among these metrics, the CTD used W_{XYA} as the inference score [19]. Before explaining W_{XYA}, we introduce the two common neighbor statistics, P_1 and P_2. In the chemical-gene-disease network, P_1 considers the number of common neighbor genes and the degree of two nodes (chemical and disease), and P_2 takes into account the degrees of common neighbor genes. According to [26], P_1 and P_2 are calculated as follows:

$$P_1(m|N, n_X, n_Y) = \frac{\binom{N}{m}\binom{N-m}{n_X-m}\binom{N-n_X}{n_Y-m}}{\binom{N}{n_X}\binom{N}{n_Y}} \quad (5)$$

$$P_2(X \text{ and } Y \text{ share } A|N) = \prod_{i\in A} \frac{n_i(n_i-1)}{N(N-1)}, \quad (6)$$

where n_X and n_Y are the node degrees of chemical X and disease Y, respectively, in a chemical-gene-disease interaction network, N is the total number of entities, m is the number of mutual neighboring nodes (genes), $A = \{Z_1,\ldots, Z_i,\ldots, Z_m\}$ is the set of common neighbor genes that connect the chemical and disease, and n_i is the number of edges of the gene Z_i in the set A.

W_{XYA} considers both the number of common genes and the connectivity among the chemical, disease, and gene. W_{XYA} [19] is the weighted product of a $\log_{10}$-transformed form of P_1 and P_2, defined as follows:

$$W_{XYA} = -(w_1 \log_{10}(P_1) + w_2 \log_{10}(P_2)) \quad (7)$$

where

$$w_1 = w_2 = (1 - \frac{e}{2e^m}).$$

To compare the reliability of the chemical-disease relations inferred using TransE with those inferred using the CTD, we constructed a chemical-gene-disease interaction network using chemical-gene, chemical-disease, and disease-gene relations corresponding to the training and validation datasets in Table 5. This network comprised 7854 chemicals, 3043 diseases, and 21,118 genes. We first applied the statistical model in CTD to this network. Thereafter, we determined the inference scores (W_{XYA}) for all pairs among 250 chemicals and 3043 diseases in the interaction network and sorted the pairs by their inference scores. These 250 test chemicals are same as those listed in Table 5. As shown in Fig 2, we already determined whether the sorted chemical-disease pairs occurred in the 71,216 remaining chemical-disease triplets for

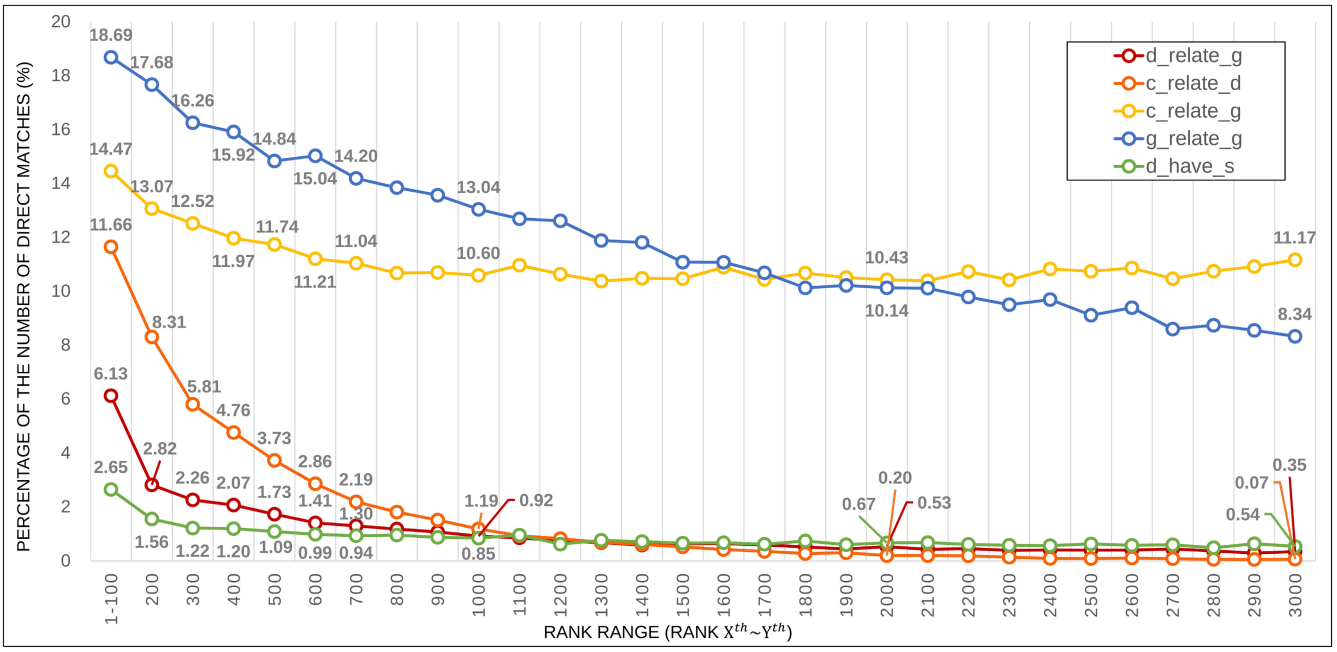

FIGURE 2. Distributions of the number of direct matches in accordance with the change in each rank. The *x*-axis represents each rank range between the top X to top Y, and the *y*-axis indicates the percentage of the number of direct matches in each rank range.

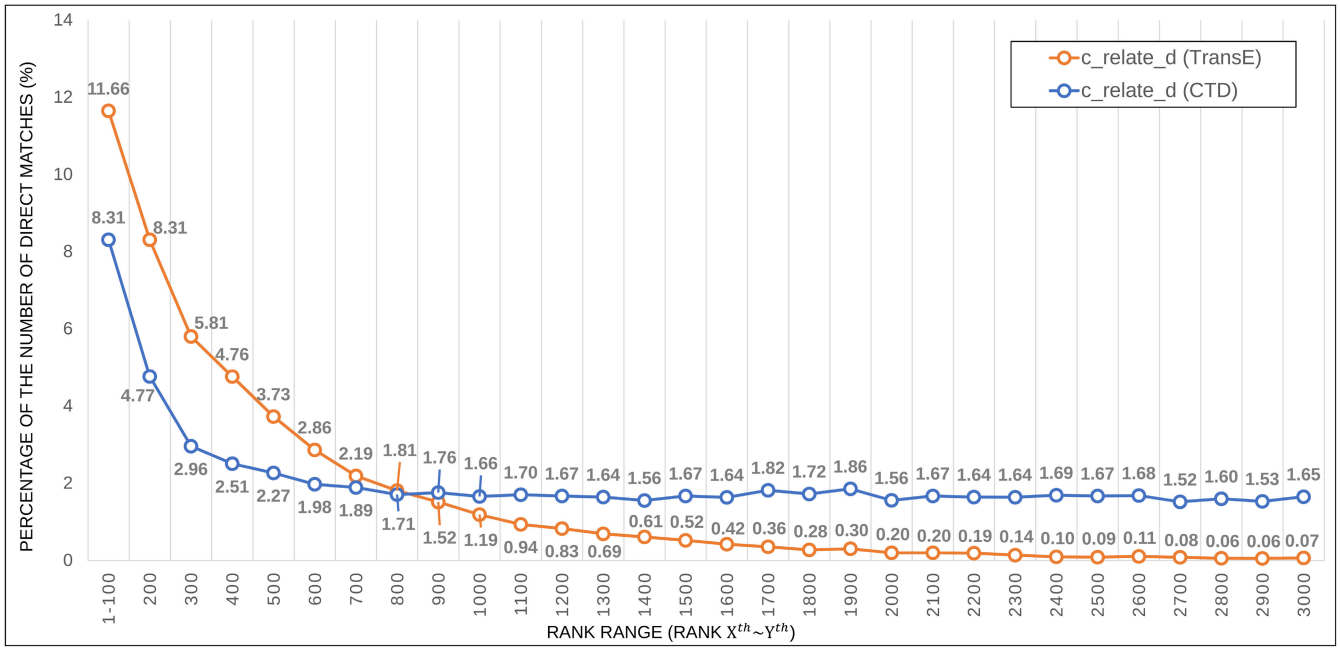

FIGURE 3. Distributions of the number of direct matches for inferred chemical-disease relations from the TransE (orange) and the Comparative Toxicogenomics Database statistical model (blue). The values on each bar represent the average number of direct matches in each rank range.

each rank range. Thus, we compared these results with those of the CTD.

As shown in the blue bar in Fig 3, when applying the statistical model of the CTD, only 8.31% of the inferred relations on average appeared in the remaining triplets in the rank range between the top 1 to 100. In contrast, an average of 11.66% of the chemical-disease relations inferred by TransE occurred in the remaining triplets. Moreover, the orange bar (TransE) tended to display a better performance than the blue bar (CTD statistical model) in the higher rank range (from rank 1 to rank 800). These results show that the relations inferred using TransE are more reliable than those inferred using the CTD statistical model. Herein, we speculated why TransE outperformed the method used in the CTD. The statistical model used in the CTD is basically based on indirect relationships between chemicals and diseases through genes to

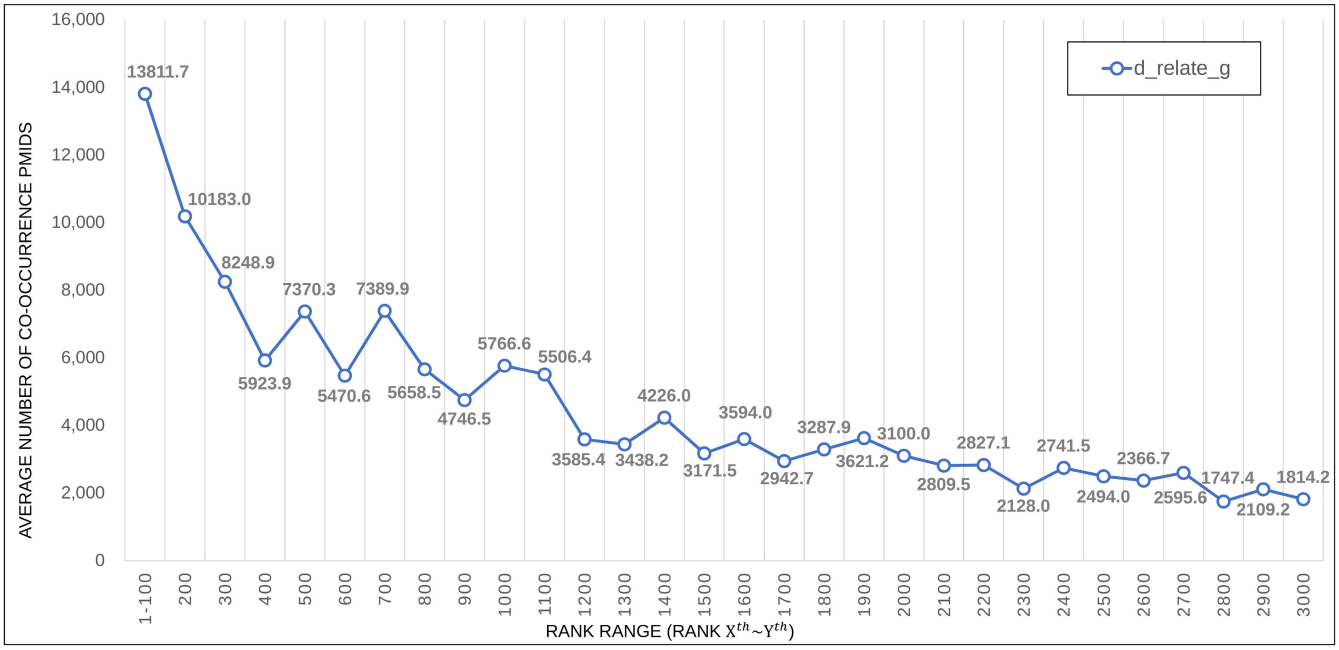

FIGURE 4. Graph explaining the distribution trends of the average number of co-occurrence PMIDs in accordance with each rank range. The *x*-axis is each rank range between top X to top Y and the *y*-axis means the average number of co-occurrence PMIDs in each rank range.

calculate their inference scores. In this process, many candidate chemical-disease relation pairs can be produced by hub genes. However, this gene-based approach generated many false-positives because the activation and regulation of drugs and genes differ depending on the type of disease. On the other hand, TransE learns vector representations of knowledge graphs in the low-dimensional vector space without the assumption of indirect relations between chemicals and diseases and determines the plausibility of a certain knowledge in the graph through algebraic operations. We assume that this difference may result in better performance of TransE.

D. EVALUATION OF INFERRED DISEASE-GENE RELATIONS ON THE BASIS OF THE NUMBER OF CO-OCCURRENCE ABSTRACTS

We hypothesized that two entities in highly ranked relations may appear in a greater number of abstracts. Thus, for 20 commonly studied diseases [9], including Alzheimer's disease and diabetic neuropathies, we inferred 3000 genes for each disease using TransE and determined the average number of abstracts wherein the inferred gene and disease are mentioned together.

As shown in Fig 4, the number of co-occurrences between the 20 diseases and corresponding inferred genes in each rank range tended to be markedly higher in the higher rank range. For example, in an average of 13,811.7 abstracts, the test diseases ranking in the top 100 co-occurred. However, the number of co-occurrence abstracts decreased at lower rank positions.

Furthermore, we determined Pearson's correlation coefficients between the ranks of genes and the number of co-occurrence abstracts. As shown in Table 6, all diseases except for "asthma" showed a significant negative relation (p-value < 0.05), indicating that higher ranked disease-gene relations tend to appear in more co-occurrence abstracts. These results support our hypothesis that highly ranked relations are more reliable than lower ranked relations.

TABLE 6. Statistics of correlation analysis for the 20 diseases.

Disease Name	Correlation coefficient value	P-value corresponding to correlation value
Alzheimer disease	-0.5529067	0.001531
Arthritis	-0.6998988	1.671e-05
Atherosclerosis	-0.696253	1.93e-05
Diabetes mellitus	-0.5712624	0.0009765
Stomach ulcer	-0.725465	5.737e-06
Hepatitis	-0.6850568	2.961e-05
Hepatocellular carcinoma	-0.7096213	1.128e-05
Hypertension	-0.5989699	0.0004702
Kidney disease	-0.684566	3.016e-05
Leukemia	-0.7329713	4.098e-06
Liver disease	-0.6779284	3.853e-05
Malaria	-0.5370682	0.002212
Melanoma	-0.6787719	3.736e-05
Obesity	-0.384613	0.03585
Parkinson disease	-0.765561	8.271e-07
Prostatic neoplasms	-0.6701133	5.1e-05
Rheumatoid arthritis	-0.6402863	0.0001385
Asthma	-0.3389832	0.06688
Diabetic neuropathy	-0.5914773	0.0005767
Dementia	-0.564551	0.001154

V. CONCLUSION AND FUTURE PROSPECTS

In this study, we constructed a biomedical knowledge base using data from well-known, publicly available databases. We compared the performance of the representation learning

models (TransE, PTransE, TransR, and TransH) based on the biomedical knowledge base. The present results show that TransE outperformed the other representation learning models despite being originally designed to overcome the limitations of TransE. Thus, we used TransE to infer novel biomedical relations among chemicals, genes, diseases, and symptoms. Thereafter, we conducted several experiments to verify the reliability of these inferred relations. In the inferred relations, higher ranked relations displayed a higher reliability than the lower ranked positions. Furthermore, the present results show that the relations inferred using TransE are more reliable than those inferred using the statistical model used in the CTD. Finally, our results show that higher ranked relations tend to be present in more co-occurrence PubMed abstracts. These findings indicate that the representation learning model is useful to infer new biomedical relations.

In future studies, we intend to develop a deep learning-based representation learning model that suitable for the present biomedical knowledge base because the performance of existing models with the biomedical data was not as satisfactory as that during the application of these models to general purpose KBs including FreeBase and WordNet.

APPENDIX

As shown in Table 2, we randomly split the biomedical data extracted from public databases into training, development, and test datasets. More specifically, we randomly picked 250 test triplets for each relation type, resulting in a total of 1250 test triplets. Thereafter, a total of 2500 development triplets (500 triplets per relation type) were randomly selected, and the rest of data were considered as the training triplets. Thus, we used them to compare the performance of each representation learning model as shown in Table 3. In this experiment, a k-fold cross-validation was not used for evaluating the performance of the models due to computational resource limitations. Unlike other KBs such as FreeBase and WordNet, our biomedical datasets consist of a total of 3,273,215 triplets (3,269,465 training + 2500 development + 1250 test triplets), which are very huge in size.

Here, we measured the training time of each representation learning model based on our biomedical datasets. TransE, which is implemented using Python and the Theano library, took 14 hours and 30 minutes. TransH took 8 hours and 33 minutes. TransR took 105 hours and 30 minutes. PTransE(ADD) took 16 hours and 40 minutes. PTransE(MUL) took 22 hours and 20 minutes. PTransE(RNN) took about 9 days. Note that all models except TransE were implemented in C++ language. Since training TransR and PTransE took a lot of time we did not use a k-fold cross-validation. All computational times have been measured on a Linux server with the following configuration: Intel Core i9-7900X CPU 3.3 GHz, NVIDIA Titan Xp CUDA GPU 12 GB GDDR5, 64 GB RAM DDR4.

REFERENCES

[1] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1156–1165.

[2] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[3] Y. Lin, Z. Liu, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," in *Proc. EMNLP*, 2015, pp. 705–714.

[4] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. 29th AAAI Conf. Artif. Intell. (AAAI)*, 2015, pp. 2181–2187.

[5] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1112–1119.

[6] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2007.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.

[8] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[9] W. Choi, C. H. Choi, Y. R. Kim, S. J. Kim, C. S. Na, and H. Lee, "HerDing: Herb recommendation system to treat diseases using genes and chemicals," *Database*, vol. 2016, pp. 1–7, Mar. 2016, Art. no. baw011, doi: 10.1093/database/baw011.

[10] C. J. Mattingly, M. C. Rosenstein, G. T. Colby, J. N. Forrest, and J. L. Boyer, "The comparative toxicogenomics database (CTD): A resource for comparative toxicological studies," *J. Exp. Zool. A, Comparative Biol.*, vol. 305, no. 9, pp. 689–692, 2006.

[11] R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen, and T. Shi, "TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D1089–D1095, Jan. 2013, doi: 10.1093/nar/gks1100.

[12] N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. I. Stein, I. Bahir, F. Belinky, C. P. Morrey, M. Safran, and D. Lancet, "MalaCards: An integrated compendium for diseases and their annotation," *Database*, vol. 2013, p. bat018, Apr. 2013, doi: 10.1093/database/bat018.

[13] D. R. Swanson, "Fish oil, raynaud's syndrome, and undiscovered public knowledge," *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[14] C. W. Tung, "ChemDIS: A chemical-disease inference system based on chemical-protein interactions," *J. Cheminform.*, vol. 7, no. 25, pp. 1–7, Jun. 2015, doi: 10.1186/s13321-015-0077-3.

[15] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: Interaction networks of chemicals and proteins," *Nucleic Acids Res.*, vol. 36, no. suppl 1, pp. D684–D688, Jan. 2008, doi: 10.1093/nar/gkm795.

[16] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, and L. M. Schriml, "Disease Ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1071–D1078, Jan. 2015, doi: 10.1093/nar/gku1011.

[17] P. Du, G. Feng, J. Flatow, J. Song, M. Holko, W. A. Kibbe, and S. M. Lin, "From disease ontology to disease-ontology lite: Statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations," *Bioinformatics*, vol. 25, no. 12, pp. i63–i68, Jun. 2009, doi: 10.1093/bioinformatics/btp193.

[18] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. suppl 1, pp. D535–D539, Jan. 2006, doi: 10.1093/nar/gkj109.

[19] B. L. King, A. P. Davis, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database," *PLoS ONE*, vol. 7, no. 11, p. e46524, Nov. 2012, doi: 10.1371/journal.pone.0046524.

[20] S. Nelson, M. Schopen, A. Savage, J. Schulman, and N. Arluk, "The MeSH translation maintenance system: Structure, interface design, and implementation," *Stud. Health Technol. Inf.*, vol. 107, no. 1, pp. 67–69, 2004.

[21] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, no. suppl 1, pp. D514–D517, Jan. 2005, doi: 10.1093/nar/gki033.
[22] S. Ayme, A. Rath, and B. Bellet, "WHO international classication of diseases (ICD) revision process: Incorporating rare diseases into the classication scheme: State of art," *Orphanet J. Rare Diseases*, vol. 5, no. suppl 1, p. P1, Oct. 2010, doi: 10.1186/1750-1172-5-S1-P1.
[23] H. Xiao, M. Huang, Y. Hao, and X. Zhu, "TransA: An adaptive approach for knowledge graph embedding," Sep. 2015, *arXiv:1509.05490*. [Online]. Available: https://arxiv.org/abs/1509.05490
[24] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "A semantic matching energy function for learning with multi-relational data," *Mach. Learn.*, vol. 94, no. 2, pp. 233–259, 2014.
[25] B. Baldwin and B. Carpenter. *LingPipe*. Accessed: Jan. 19, 2015. [Online]. Available: http://www.alias-i.com/lingpipe/
[26] H. Li and S. Liang, "Local network topology in human protein interaction data predicts functional association," *PLoS ONE*, vol. 4, no. 7, p. e6410, Jul. 2009, doi: 10.1371/journal.pone.0006410.

WONJUN CHOI was born in Seoul, South Korea, in 1988. He received the B.S. degree in computer engineering from Korea Aerospace University, in 2013, and the M.S. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2015, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science. His research interests include data mining, recommendation systems and corpus construction.

HYUNJU LEE received the B.S. degree in computer science from the Korea Institute of Science and Technology, South Korea, in 1997, the M.S. degree in computer engineering from Seoul National University, South Korea, in 1999, and the Ph.D. degree in computer science from the University of Southern California, USA, in 2006. From 2006 to 2007, she was a Postdoctoral Research Fellow with the Brigham Women's Hospital, Harvard Medical School. Since 2007, she has been with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. Her research interests include machine learning, natural language processing, and bioinformatics.