Representation Learning for Measuring Entity Relatedness with Rich Information

Yu Zhao[1], **Zhiyuan Liu**[1,2*], **Maosong Sun**[1,2]
[1] Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology, Tsinghua University, Beijing, China
[2] Jiangsu Collaborative Innovation Center for Language Ability,
Jiangsu Normal University, Xuzhou 221009 China
zhaoyu.thu@gmail.com, {liuzy, sms}@tsinghua.edu.cn

Abstract

Incorporating multiple types of relational information from heterogeneous networks has been proved effective in data mining. Although Wikipedia is one of the most famous heterogeneous network, previous works of semantic analysis on Wikipedia are mostly limited on single type of relations. In this paper, we aim at incorporating multiple types of relations to measure the semantic relatedness between Wikipedia entities. We propose a framework of coordinate matrix factorization to construct low-dimensional continuous representation for entities, categories and words in the same semantic space. We formulate this task as the completion of a sparse entity-entity association matrix, in which each entry quantifies the strength of relatedness between corresponding entities. We evaluate our model on the task of judging pair-wise word similarity. Experiment result shows that our model outperforms both traditional entity relatedness algorithms and other representation learning models.

1 Introduction

Heterogeneous text-rich networks, such as social network sites, question answering communities and knowledge graphs, consist of various types of entities as well as large amounts of textual documents [Deng *et al.*, 2011]. They have been popularly used in the research of information retrieval and data mining [Sun and Han, 2012]. Specifically, as the largest online encyclopedia in existence, Wikipedia has been used for knowledge acquisition and semantic analysis for years [Medelyan *et al.*, 2009]. The rich information contained in both the content of articles and the structure of linkage network makes it of great worth for incorporating cross-domain knowledge and interpreting semantic relationships.

In this paper, we aim at measuring semantic relatedness between Wikipedia entities. Former researches usually transform this task into measuring semantic similarity between word senses. Most of existing semantic relatedness measures can be divided into the following three typical types:

Text-theoretic measures. These measures build word representations directly from statistical information in text corpus, since related entities tends to be described by similar words. Traditional methods mainly focus on counting the occurrences of words appearing in each Wikipedia article to construct a high-dimensional semantic space [Salton *et al.*, 1975; Gabrilovich and Markovitch, 2009]. In recent years, low-dimensional word representations have achieved great success in many natural language processing tasks [Collobert *et al.*, 2011]. They represent the semantic meaning of word by a vector of latent continuous features.

Graph-theoretic measures. These measures usually take advantage of the hyperlink structure of entity graph. The first strategy is to find the shortest path between given entities [Leacock and Chodorow, 1998]. For instance, if the length of shortest path between *stock* and *petroleum* is short, they tend to be treated as related. The second strategy is to compare the neighbors of given entities [Witten and Milne, 2008]. For instance, if the number of common neighbors of *stock* and *petroleum* is large, they tend to be treated as related.

Information-theoretic measures. These methods usually takes advantage of the Wikipedia Category taxonomy. It mainly focuses on calculating the information content of each category [Strube and Ponzetto, 2006]. For instance, the information content value of category *Fuels* is lower than *Finance* because the former has more parent nodes. Relatedness between *stock* and *petroleum* is measured by the information content of their lowest common subsumer *Economies*.

However, there are two main disadvantages of these measures in practice:

Sparsity. The interactions between entities are relatively sparse compared with the total number of entity-entity pairs. Some of the description articles have even no hyperlinks to other entities. The structure of category taxonomy is even more sparse. In addition, the length of paths between two entities is always limited to a small number. It indicates that path-based measures are relatively less discriminative.

Homogeneity. Most of these measures do not consider the interactions between vertices of different types in the network. It is difficult for them to simultaneously incorporate multiple types of relations, such as entity-word and entity-category relations, which reduces their scalability and extensibility.

In this paper, we introduce a matrix factorization frame-

*Corresponding author: Zhiyuan Liu (liuzy@tsinghua.edu.cn).

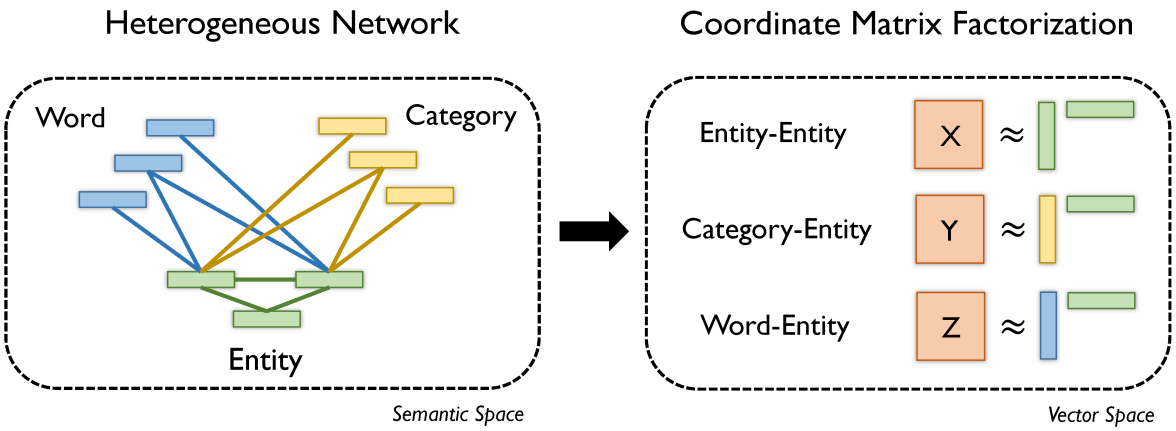

Figure 1: The framework of utilizing multiple types of relation in heterogeneous network for representation learning via coordinate matrix factorization. In the case of analyzing Wikipedia, the vector representations for entities, categories and words are constructed simultaneously

.

work to overcome these problems. We formulate the task of measuring entity relatedness as the completion of an entity-entity association matrix, where each entry in the matrix quantifies the strength of relatedness between corresponding entities. In addition, we exploit the random walk process on graph to obtain large amounts of entity sequences, which can capture the co-occurrences between disconnected but related vertices.

In order to incorporate multiple types of relation, we propose to generate a prior association matrix for each component pair, such as entity-entity matrix, category-entity matrix and word-entity matrix. Each entry in the association matrices is modeled as the inner product of two low-dimensional component vectors, as demonstrated in Figure 1. A low-dimensional latent vector space is found for all three components of Wikipedia: entities, categories and words. The representations for all types of components are synchronously constructed in the same semantic space.

We present experiments on word similarity tasks to evaluate the performance of our model. In accordance with the original mission of measuring entities relatedness, we need to verify that each word appearing in the dataset correspond to an entity in Wikipedia. Experiment result shows that our model outperforms both traditional entity relatedness algorithms and other representation learning models.

The key contribution of our work lies in:

- We introduce the coordinate matrix factorization model to measure semantic relatedness between Wikipedia entities. It naturally overcomes the sparsity problem and ensures the relatedness value to be discriminative.
- We present a flexible framework of incorporating multiple types of relations, which can be easily extended by introducing more association matrices in consideration.
- We conducted intensive experiments on word similarity task. Results show that our model achieves a better performance compared with previous methods.

2 Problem Definition

In this section, we formally define the problem of learning entity representation and introduce the notations used in our model.

As a text-rich heterogeneous network with different types of components, Wikipedia can be denoted as a graph $G = (\chi, \xi)$ where χ is a set of components and ξ is a set of edges between components. Specifically, Wikipedia consists of three types of components, including an entity set E, a category set C and a vocabulary set W, hence $\chi = E \cup C \cup W$. Meanwhile, edges in ξ can be divided into three types of relations, including internal hyperlinks between entities $L_E = \{(e_i, e_j)|e_i, e_j \in E\}$, category labels for each entity $L_C = \{(c_i, e_j)|c_i \in C, e_j \in E\}$, and the occurrences of words in the description article of each entity $L_W = \{(w_i, e_j)|w_i \in W, e_j \in E\}$, hence $\xi = L_E \cup L_C \cup L_W$.

We aim at discovering a joint latent semantic space H for all three components with common dimensionality K. We define three low-rank matrices $\mathbf{E_H} \in \mathbb{R}^{|E|\times K}$, $\mathbf{C_H} \in \mathbb{R}^{|C|\times K}$ and $\mathbf{W_H} \in \mathbb{R}^{|W|\times K}$ referring to ensembles of vectors. Each entity $e_i \in E$ is associated with a continuous feature vector $\mathbf{E_{H_i}} \in \mathbb{R}^K$, namely the i-th row vector in $\mathbf{E_H}$. Similarly, each category $c_j \in C$ and each word $w_l \in W$ are associated with row vectors $\mathbf{C_{H_j}} \in \mathbb{R}^K$ and $\mathbf{W_{H_l}} \in \mathbb{R}^K$ respectively.

Once we obtain a vector $\mathbf{E_{H_i}}$ for each entity e_i, the semantic relatedness $sr(e_i, e_j)$ of two entities e_i and e_j can be computed by the cosine similarity of corresponding vectors, which is formulated as:

$$sr(e_i, e_j) = \frac{\sum_{l=1}^{K} E_{H_{il}} \cdot E_{H_{jl}}}{\sqrt{\sum_{l=1}^{K} E_{H_{il}}^2} \cdot \sqrt{\sum_{l=1}^{K} E_{H_{jl}}^2}} \quad (1)$$

Our goal is to learn an representation matrix $\mathbf{E_H}$ that approximates the non-zero entries in an entity-entity coefficient matrix $\mathbf{X} \in \mathbb{R}^{|E|\times|E|}$. In the learning procedure, the representation of categories $\mathbf{C_H}$ and words $\mathbf{W_H}$ are obtained at the

same time. In Section 5 we will show that incorporating category and word representations contributes to the performance of learning entity representation.

3 Coordinate Matrix Factorization

In our model, the procedure of learning entity representation matrix $\mathbf{E_H} \in \mathbb{R}^{|E|\times K}$ is formulated as the estimation of an entity-entity coefficient matrix $\mathbf{X} \in \mathbb{R}^{|E|\times|E|}$. Each entry X_{ij} in matrix $\mathbf{X}$ can be generated using the entity linkage subset $L_E \subseteq \xi$.

The quality of the estimation is evaluated using the Non-Zero Square Loss (NZSL), which is equivalent to minimizing the following objective function with weighted L_2 regularization term:

$$\sum_{(i,j)\in N_{\mathbf{X}}} (\mathbf{X}_{ij} - \mathbf{E}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{L}_j}^\top)^2 + \lambda(\|\mathbf{E_H}\|_F^2 + \|\mathbf{E_L}\|_F^2), \quad (2)$$

where $\mathbf{E_L} \in \mathbb{R}^{|E|\times K}$ is the ensemble of latent constraint vectors for each entity, $N_{\mathbf{X}} = \{(i,j) : X_{ij} \neq 0\}$ refers to the non zero entries in $\mathbf{X}$, λ is the regularization factor and $\|\cdot\|_F^2$ is the Frobenius norm. We will introduce the strategy of generating $\mathbf{X}$ in Section 4 in detail.

To incorporate the relation between entities and words, or between entities and categories, we construct two coefficient matrix for each component pair respectively: $\mathbf{Y} \in \mathbb{R}^{|C|\times|E|}$ for entity-category relations, and $\mathbf{Z} \in \mathbb{R}^{|W|\times|E|}$ for entity-word relations. In our model, $\mathbf{Y}$ is estimated by the inner product of entity representation matrix $\mathbf{E_H}$ and category representation matrix $\mathbf{E_C}$, while $\mathbf{Z}$ is estimated by the inner product of $\mathbf{E_H}$ and word representation matrix $\mathbf{E_W}$. Then the overall loss function is formulated as:

$$\begin{aligned} J = &\sum_{(i,j)\in N_{\mathbf{X}}} (\mathbf{X}_{ij} - \mathbf{E}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{L}_j}^\top)^2 + \sum_{(i,j)\in N_{\mathbf{Y}}} (\mathbf{Y}_{ij} - \mathbf{C}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{H}_j}^\top)^2 \\ &+ \sum_{(i,j)\in N_{\mathbf{Z}}} (\mathbf{Z}_{ij} - \mathbf{W}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{H}_j}^\top)^2 + \lambda(\|\mathbf{E_H}\|_F^2 + \|\mathbf{E_L}\|_F^2) \\ &+ \gamma\|\mathbf{C_H}\|_F^2 + \delta\|\mathbf{W_H}\|_F^2, \end{aligned} \quad (3)$$

where $N_{\mathbf{Y}}$ and $N_{\mathbf{Z}}$ refer to the non-zero entries in $\mathbf{Y}$ and $\mathbf{Z}$, with λ, γ, δ are the regularization factors. In the experiment we use $\lambda = 0.01$ and $\gamma = \delta = 0.005$.

We adopt a coordinate stochastic gradient descent scheme to train this model. After initializing each representation matrices $\mathbf{E_H}$, $\mathbf{E_L}$, $\mathbf{C_H}$ and $\mathbf{W_H}$, we iteratively update one of these matrices with all others fixed. The training procedure stops when it reaches an upper bound of iteration times or achieves an optimal rate of convergence.

During the learning process, we only consider the local loss over one entry in the training set at a time. Take the matrix $\mathbf{X}$ for example, when observing a non-zero entry $(i,j) \in N_{\mathbf{X}}$, we only update the elements $E_{H_{ik}}$ and $E_{L_{kj}}$ in related local vectors $\mathbf{E}_{\mathbf{H}_i}$ and $\mathbf{E}_{\mathbf{L}_j}$ by enumerating the position k. The derivatives of the objective function with respect to each entry are as follows:

$$\begin{aligned} \frac{\partial J_{\mathbf{X}}}{\partial E_{L_{ik}}} &= -2(\mathbf{X}_{ij} - \mathbf{E}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{L}_j}^\top)E_{H_{kj}} + 2\lambda\frac{E_{L_{ik}}}{N_{\mathbf{X}_{i*}}} \\ \frac{\partial J_{\mathbf{X}}}{\partial E_{H_{kj}}} &= -2(\mathbf{X}_{ij} - \mathbf{E}_{\mathbf{H}_i}\mathbf{E}_{\mathbf{L}_j}^\top)E_{L_{ik}} + 2\lambda\frac{E_{H_{kj}}}{N_{\mathbf{X}_{*j}}}, \end{aligned} \quad (4)$$

where $N_{\mathbf{X}_{i*}}$ is the number of non-zero entries that appear in row i of matrix $\mathbf{X}$, and $N_{\mathbf{X}_{*j}}$ is the number of non-zero entries in column j of matrix $\mathbf{X}$. At each step of coordinate stochastic gradient descent, we randomly select a sample of $(i,j) \in N_{\mathbf{X}}$ and make the following updates for each k:

$$\begin{aligned} \triangledown E_{L_{ik}} &= E_{L_{ik}} - \eta \cdot \frac{\partial J_{\mathbf{X}}}{\partial E_{L_{ik}}} \\ \triangledown E_{H_{kj}} &= E_{H_{kj}} - \eta \cdot \frac{\partial J_{\mathbf{X}}}{\partial E_{H_{kj}}}, \end{aligned} \quad (5)$$

where η is the learning rate. In the experiment we initialize η with 0.01 and linearly decrease it after each iteration.

In a similar manner, we can derive an alternative update for $\partial J_{\mathbf{Y}}/\partial \mathbf{E_H}$, $\partial J_{\mathbf{Y}}/\partial \mathbf{C_H}$ given an entry in $\mathbf{Y}$, and $\partial J_{\mathbf{Z}}/\partial \mathbf{E_H}$, $\partial J_{\mathbf{Z}}/\partial \mathbf{W_H}$ given an entry in $\mathbf{Z}$.

4 Generation of Coefficient Matrices

In the scheme of coordinate matrix factorization, the generation process of the coefficient matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ significantly influences the performance. It is very challenging to directly learn the representation from observed edges in $\xi \in G$ due to its sparsity. In this paper, we make use of Random Walk and Positive Pointwise Mutual Information (PPMI) to preset the non-zero entries in the coefficient matrices to overcome the sparsity problem.

4.1 Random Walk

Random walk has been used in the task of learning representations of vertices in a network, such as DeepWalk [Perozzi *et al.*, 2014]. Given a vertex v in a network, a *random walk* from root v is a stochastic process to generate an output path $p = (v, v_1, v_2, \ldots, v_n)$, where v_{k+1} is a vertex chosen at random from the neighbors of vertex v_k. The main idea of exploiting random walk on graph is to obtain large amounts of vertex sequences, in order to capture the co-occurrences between disconnected vertices.

As mentioned in Section 2, we consider Wikipedia as graph $G = (\chi, \xi)$, where $\chi = E \cup C \cup W$ and $\xi = L_E \cup L_C \cup L_W$. To distinguish multiple types of relations for different coefficient matrices, we split G into three subgraphs: $G_E = (E, L_E)$, $G_C = (E \cup C, L_C)$ and $G_W = (E \cup W, L_W)$. For each subgraph, we enumerate all of its vertices as the root of random walk paths. We set the length of a random walks to be fixed with m, and the number of output paths per vertex to be fixed with t. Hence, we obtain $|E| \cdot t$ random walks from G_E, $(|E| + |C|) \cdot t$ random walks from G_C and $(|E| + |W|) \cdot t$ random walks from G_W. For each subgraph, we collect all random walk paths into a document D and treat it as a corpus of special sentences. For example, the document D_E of subgraph G_E is defined as:

$$D_E = \{(e_1, e_2, \ldots, e_m) | e_i \in E, (e_k, e_{k+1}) \in L_E\}. \quad (6)$$

Notice that both G_C and G_W are bipartite graphs, so that each bigram of vertices (e_k, v_{k+1}) or (v_k, e_{k+1}) in documents D_C and D_W contains one and only one entity. In the experiment we set $t = 50$ and $m = 10$.

4.2 Positive Pointwise Mutual Information

Recently, word representation models have been shown to perform well in various NLP tasks. One of the most famous training method is Skip-Gram model (SGM) [Mikolov *et al.*, 2013]. Given a corpus of word sequences, the objective of SGM with negative sampling is to maximize:

$$\log \sigma(\vec{w} \cdot \vec{c}_w) + \sum_{i=1}^{k} E_{w_i \sim Pn(w)}[\log \sigma(-\vec{w} \cdot \vec{c}_{w_i})] \qquad (7)$$

where $\vec{w}$ and $\vec{c}_w$ are vector representations for word w and its surrounding context c_w, $Pn(w)$ is a set of negative word samples, and $\sigma(\cdot)$ is the sigmoid function. Since we have collected documents D_E, D_C and D_W for subgraphs in Wikipedia, it is intuitive to replace $\vec{w}$ and $\vec{c}_w$ by representations of vertices in these documents, such as $\mathbf{E}_{\mathbf{H}_i}$ and $\mathbf{E}_{\mathbf{L}_j}$. The difference is that word representation models make use of local context information within words, while we exploit relations between different types of vertices in Wikipedia.

In fact, Levy and Goldberg [2014] have proved that the objective of SGM is implicitly factorizing an coefficient matrix of words and contexts. We adopt their derivation in our task for purpose of generating coefficient matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$. For example, each entry in $\mathbf{X}$ is measured by:

$$\mathbf{E}_{\mathbf{H}_i} \cdot \mathbf{E}_{\mathbf{L}_j} = \log \left(\frac{\#(e_i, e_j) \cdot |D_E|}{\#(e_i) \cdot \#(e_j)} \right) - \log k \qquad (8)$$

where $\#(e_i, e_j)$ denotes the number of times e_i and e_j co-occur in the sentences of D_E, $\#(e_i) = \sum_l \#(e_i, e_l)$ and $\#(e_j) = \sum_l \#(e_l, e_j)$ denote the number of times e_i and e_j appear in D_E.

Notice that when $k = 0$, the Equation 8 is equivalent to PMI because the probability distribution of an entity e can be interpreted as $P(e) = \frac{\#(e)}{|D_E|}$ and the probability of co-occurrence is $P(e_i, e_j) = \frac{\#(e_i, e_j)}{|D_E|}$.

In our model, we omit those entries with negative PMI values, which tend to indicate that the corresponding vertices are not related. We use the Positive PMI value to generate the coefficient matrices:

$$\mathbf{X}_{ij} = \text{PPMI}(e_i, e_j) = \max \left(\log \frac{P(e_i, e_j)}{P(e_i) \cdot P(e_j)}, 0 \right) \qquad (9)$$

Similarly, the entries in coefficient matrices $\mathbf{Y}$ and $\mathbf{Z}$ are approximated by PPMI(c_i, e_j) and PPMI(w_i, e_j) respectively. We finally obtain the representation for entities by implementing the coordinate matrix factorization introduced above in Section 3.

5 Experiments

In this paper, we evaluate our methodology by judging the similarity between words. We first introduce the dataset used in the experiment. Then we demonstrate that our model performs the best compared with all baseline methods.

5.1 Dataset

In accordance with the task of measuring relatedness of entities instead of words, we need to verify that each word sense appearing in the dataset corresponds to an entity in Wikipedia. We select the word similarity dataset Words-240 [Xiang *et al.*, 2014][1]. This dataset contains 240 pairs of Chinese words, each of which is labeled by 20 annotators, which ensures its reliability. Human ratings of pairwise similarity ranges from 0 to 10, where a high number indicates higher similarity.

To confirm that this dataset can be used for evaluating entity representations, we collect the whole Chinese Wikipedia corpus online. It consists of 291,475 entities, 156,601 categories and a vocabulary of 365,650 unique words after word segmentation and removing low-frequent words. In the dataset, we find 9 word pairs that contain words not related to Wikipedia entities. Hence, in this paper we only reserve the rest 231 word pairs for evaluation. We use the same segmented Chinese Wikipedia corpus for all methods, which ensures fair comparison.

In our experiment, we compute the semantic relatedness with entity representation using Equation 1. The result is evaluated against the human similarity ratings using Spearman's ρ correlation coefficient.

5.2 Baseline Methods

In this paper, we compare our model against the following baselines, including information-theoretic models, graph-theoretic models and text-theoretic models:

- **Information Content** (IC): Resnik [1995] computes the semantic relatedness of two vertices by the information content of the concepts that subsume them in the taxonomy, which is formulated as:

$$sr(e_1, e_2) = \max_{\substack{c_1, c_2, \\ c \in S(c_1, c_2)}} [-\log p(c)] \qquad (10)$$

  where c_1, c_2 are category labels of w_1, w_2 respectively, $S(c_1, c_2)$ is their common subsumers in category taxonomy C, and the occurrence probability is $p(c) = \frac{\#(C)}{D_C}$.

- **Shortest Path** (SP): Leacock and Chodorow [1998] uses the length of the shortest path between two vertices to compute semantic relatedness, which is formulated as:

$$sr(e_1, e_2) = -\log \frac{length(e_1, e_2)}{2 \times depth} \qquad (11)$$

  where $depth$ stands for the length of the longest path in the undirected entity subgraph G_E.

- **Information Distance** (ID): If we treat the Wikipedia entity linkage graph G_E as a directed graph, the shortest path between e_1 and e_2 may not be reversible. Hence we also compare with the information distance based on

[1] We did not select the famous English dataset wordsim-353 because a higher proportion of word senses in it are not entities in Wikipedia, such as defeating, focus, delay, attempt, recommendation, stupid, etc.

the conditional Kolmogorov complexity [Li and Vitányi, 2009]:

$$sr(e_1, e_2) = -\log \frac{\max\{length(e_1, e_2), length(e_2, e_1)\}}{2 \times depth} \tag{12}$$

- **Wikipedia Link-based Measure** (WLM): Milne and Witten [2008] proposed a low-cost measure of semantic relatedness based on Wikipedia entity graph, inspired by Normalized Google Distance [Cilibrasi and Vitanyi, 2007]:

$$sr(e_1, e_2) = \frac{\log(\max(|E_1|, |E_2|)) - \log(|E_1 \cap E_2|)}{\log(|E|) - \log(\min(|E_1|, |E_2|))} \tag{13}$$

where E_1 and E_2 are the sets of all entities that link to e_1 and e_2 respectively.

- **Skip-gram Model** (SGM): Skip-Gram model [Mikolov *et al.*, 2013] utilizes textual information to capture latent word relationships. We use Wikipedia articles as training corpus to learn word vectors. Hyperlinks between entities are reserved and represented by unique tokens.
- **DeepWalk** (DW): DeepWalk [Perozzi *et al.*, 2014] learns representations of vertices in a graph with a random walk generator and language modeling. We use the combination of random walk path documents $D = D_E \cup D_C \cup D_W$ to train this model.

5.3 Experiment Results

We report the experiment results of four variations of Coordinate Matrix Factorization (CMF) models and six baselines on the dataset. In the experiment, we set the dimensionality K of the vector space to be 200. We respectively collect 182,774,540, 10,199,647 and 35,926,605 non-zero entries in the coefficient matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$.

Table 1 demonstrates the performances of different methods. As we can see, when we exploit all of the three coefficient matrices, CMF achieves a highest of 0.465 on Spearman's ρ correlation coefficient with human ratings, which is at least 0.023 higher compared with any of the six baselines. This result shows that our representation learning method can better predict the similarity between entities. Moreover, CMF performs better when more coefficient matrices are used. It indicates that incorporating multiple types of relations in Wikipedia contributes to the performance of learning entity representation. Specifically, we find that the performance of CMF with matrices $\mathbf{X}, \mathbf{Y}$ is better than that with $\mathbf{X}, \mathbf{Z}$, even if $\mathbf{Z}$ contains twice more non-zero entries than $\mathbf{Y}$. It indicates that entity-category relation is more beneficial for measuring entity relatedness.

We also notice that with the same types of relation used, CMF achieves the best performance in all circumstances. If we only use the entity-entity relation in $\mathbf{X}$, CMF achieves 0.4413 on Spearman's ρ correlation, which is higher than the results of SP, ID and WLM. If we use additional entity-category relation in $\mathbf{Y}$, CMF significantly outperforms IC.

Surprisingly, we find that CMF with entity-entity and entity-word relations in $\mathbf{X}, \mathbf{Z}$ outperforms SGM with an improvement of 0.024, even though we retain both anchor texts and referenced entities in the content of Wikipedia articles as training corpus. The reason is that SGM only captures the co-occurrences of entities and words in a limited context window, while CMF captures the relation between an entity and all words in the corresponding description article. Moreover, the number of hyperlinks in Wikipedia limited the amount of entity-word relations that SGM can access.

	Model	Relation types used	$\rho \times 100$
Baselines	IC	E-C + C-C	36.94
	SP	E-E	40.57
	ID	E-E	41.33
	WLM	E-E	41.09
	SGM	E-W + W-W	43.26
	DW	E-E + E-C + E-W	44.17
CMF	$\mathbf{X}$	E-E	44.13
	$\mathbf{X}, \mathbf{Y}$	E-E + E-C	46.38
	$\mathbf{X}, \mathbf{Z}$	E-E + E-W	45.72
	$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	E-E + E-C + E-W	**46.50**

Table 1: Spearman's ρ correlation of different models with human ratings, where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ stand for the coefficient matrices used in our Coordinate Matrix Factorization (CMF) model. We respectively denote entity, category and word as E, C and W, to present different types of relations used between vertices in Wikipedia.

Finally, we find that CMF performs better than DeepWalk using the same types of relations, which confirms that unsupervised feature learning via neural language model is approximated to matrix factorization on specific coefficient matrices. It is worth mentioning that if we strictly follow the Shifted PPMI metric [Levy and Goldberg, 2014] with the same negative-sampling parameter, DeepWalk is implicitly factorizing CMF model with coefficient matrix $\overline{\mathbf{X}} = \mathbf{E_H} \cdot \mathbf{E_H}^\top$ instead of current $\mathbf{X} = \mathbf{E_H} \cdot \mathbf{E_L}^\top$.

5.4 Computational Cost

The validation of performance implies not only effectiveness but also efficiency. Among all the baselines, IC, SP, ID and WLM are heuristic functions with no training phrase, but during testing phrase they are thousands of times slower than vector based models. SGM and DeepWalk both utilize Skip-Gram model for representation learning, which traverses all co-occurrences of entity pairs. CMF model compresses the information into association matrices, so that the time cost of each iteration is smaller. However, CMF requires larger number of iterations to converge because it optimizes an overall loss function involving the whole matrices.

In the experiment, we use a computer with eight Intel Xeon 2.00GHz processors for parallel computing. It takes three hours for Skip-Gram model to converge. For CMF model, the average time cost of an iteration is 4.6 minutes and it converges after 220 iterations. It indicates that the time cost of CMF is higher than Skip-Gram model but still feasible.

5.5 Case Study

To demonstrate the effectiveness of entity representation in CMF model, we provide an example of three entities and

Music	Machine Learning	Cookie
composer	ANN	pie
musician	computer vision	cheesecake
classical music	data mining	pudding
popular music	pattern recognition	dim sum
jazz	speech recognition	apple pie
rhythm	NLP	butter
vocal music	machine vision	flatbread
melody	face perception	wheat gluten
orchestra	randomness	cake
musical instrument	cluster analysis	biscuit roll

Table 2: Examples of top 10 nearest neighbor entities in the vector space of CMF model (translated in English).

their top 10 nearest neighbors, as shown in Table 2. We observe that semantically related entities are closer in the vector space, which validates that our entity representation is capable of measuring semantic relatedness.

In order to gain more information on the limitation of our model, we provide some examples of severe conflicts between CMF and human ratings, as can be seen in Table 3. We find that CMF has trouble in identifying implicit relations, which is usually straight forward to humans. For instance, the correlation between *movie* and *popcorn*, *physician* and *obligation* originates from human's feeling and life experience, which is hard to infer from the link structure of entity graph or from the content of articles.

Meanwhile, some word pairs are considered as related by CMF but are assigned with low human ratings, such as *future* and *prophecy*, *chemistry* and *atom*. They tend to appear in each other's description article and provides large amounts co-occurrence information. However, these entities are usually about science or philosophy, so it is more difficult for annotators to be conscious of their relevance.

6 Related Work

Traditional measures of semantic relatedness build word representation directly from statistical co-occurrences in text corpus. Words are represented in a high-dimensional semantic space depending on the bag of words representation [Salton *et al.*, 1975], but human semantic space is actually of fairly low dimension. Hence, a variety of vector-space models have been proposed to decrease the dimensionality and to interpret the latent meaning of each dimension. Latent Semantic Analysis [Landauer and Dumais, 1997] reduces the dimensionality by factorizing a word-by-document matrix. Explicit Semantic Analysis [Gabrilovich and Markovitch, 2009] interprets each dimension by concepts (i.e. entities) in Wikipedia. These methods are proved effective on constructing word representations and measuring word similarities.

With the growth of Wikipedia, more researchers have begun to focus on the structure of its article graph and category taxonomy. Ollivier and Senellart [2007] use article graph structure to find related entities. Ponzetto and Strube [2006] use category taxonomy structure to estimate semantic relatedness between entities. Chernov *et al*. [2006] also extract semantically related categories with the aid of category links.

Entity 1	Entity 2	CMF Rank	Human Rank
shower	water	228	18
physician	obligation	185	4
movie	popcorn	229	60
drinking	mouth	179	14
earth	human	210	76
food	growth	207	85
daytime	night	144	5
bed	wardrobe	40	185
future	prophecy	41	177
tiger	cat	67	201
psychology	cognition	36	158
chemistry	atom	32	149
reproduction	egg	80	188
planet	astronomer	24	119

Table 3: Most severe conflicts between human judgments and CMF decisions in rating entity similarity, ranging from rank 1 to rank 230 (translated in English).

West *et al*. [2009] aim at collecting human click statistics to help compute semantic distance.

Yeh *et al*. [2009] first manage to explore the use of multiple link types taken as a uniform graph in Wikipedia. They applied random walks on Wikipedia article graph with Personalized PageRank. Shirakawa *et al*. [2009] first introduce the idea of constructing high-dimensional concept vectors from Wikipedia category network. In contrast, neural language models can induce low-dimensional word representations [Turian *et al.*, 2010], but the latent semantic meaning of each dimension can not be interpreted. Perozzi *et al*. [2014] first take advantage of neural language models in learning vector representation for vertices in the network.

Matrix factorization maps different factors to a joint latent space of the same dimensionality. It is widely used in collaborative filtering and recommender systems [Koren *et al.*, 2009]. Levy and Goldberg *et al*. [2014] first introduce this technique for learning word representation. They proved that word representation model is implicitly factorizing a specific word-context matrix.

7 Conclusion

In this paper, we propose a coordinate matrix factorization based method for measuring semantic relatedness on Wikipedia, which incorporates multiple types of relation between entities, categories and words. We construct low-dimensional continuous representations of entities, and formulate the task of measuring entity relatedness as the completion of an originally sparse entity-entity association matrices. Experiment result shows that our model achieves better performance compared with traditional measures, neural word representations and other entity representations.

Acknowledgments

This research is supported by the 973 Program (No. 2014CB340501) and the National Natural Science Foundation of China (NSFC No. 61133012 & 61170196).

References

[Chernov *et al.*, 2006] Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. Extracting semantics relationships between wikipedia categories. *SemWiki*, 206, 2006.

[Cilibrasi and Vitanyi, 2007] Rudi L Cilibrasi and Paul MB Vitanyi. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383, 2007.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[Deng *et al.*, 2011] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1271–1279. ACM, 2011.

[Gabrilovich and Markovitch, 2009] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, pages 443–498, 2009.

[Koren *et al.*, 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[Landauer and Dumais, 1997] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[Leacock and Chodorow, 1998] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[Li and Vitányi, 2009] Ming Li and Paul MB Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2009.

[Medelyan *et al.*, 2009] Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[Ollivier and Senellart, 2007] Yann Ollivier and Pierre Senellart. Finding related pages using green measures: An illustration with wikipedia. In *AAAI*, volume 7, pages 1427–1433, 2007.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[Resnik, 1995] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[Salton *et al.*, 1975] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[Shirakawa *et al.*, 2009] Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Concept vector extraction from wikipedia category network. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*, pages 71–79. ACM, 2009.

[Strube and Ponzetto, 2006] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424, 2006.

[Sun and Han, 2012] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.

[Turian *et al.*, 2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[West *et al.*, 2009] Robert West, Joelle Pineau, and Doina Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI*, pages 1598–1603, 2009.

[Witten and Milne, 2008] I Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[Xiang *et al.*, 2014] Wang Xiang, Jia Yan, Zhou Bin, Zhaoyun Ding, and Zheng Liang. Computing semantic relatedness using chinese wikipedia links and taxonomy. In *Journal of Chinese Computer Systems*, pages 2177–2185, 2014.

[Yeh *et al.*, 2009] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics, 2009.