A Constructivist Ontology Relation Learning Method

Zhenping Xie, Liyuan Ren, Qianyi Zhan, and Yuan Liu

***Abstract*—From the perspective of philosophy, ontology relations denote ultimate semantic relations of related knowledge concepts. Beyond doubt, it is still a very difficult problem on how to automatically depict and construct ontology relations because of its high abstractness. Some latest research attempted to realize ontology relation learning by learning abstract hierarchies or similarities among knowledge concepts. Inspired by the requirements of associative semantic cognition like in the human brain, a constructivist ontology relation learning (CORL) method is put forward in this study by borrowing the idea of the constructivist learning theory. Wherein, two following points are supposed: 1) each symbol knowledge is looked as a token of representing certain abstract pattern and 2) each pattern denotes a type of relation structures on other patterns, or a directly observed event data, such as physical sensing data, natural image, sound data, text word etc. So, ontology relation could be considered as the associative support degrees from other knowledge concepts to the target concept, which reflects how one knowledge ontology can be demarcated by other knowledge concepts. Then, the knowledge network can be employed to represent an entire domain knowledge system. Meanwhile, an associative random walk mechanism (ARWM) on knowledge network can be considered to explain the semantic generative process of every document. Thus, CORL can be realized by integrating ARWM into an extended latent Dirichlet allocation (LDA) model. Some theoretical and experimental analysis are done. The corresponding results demonstrate that CORL can obtain effective associative semantic relations among concept words, and gain some novel characteristics in better representing knowledge ontology than existing methods.**

***Index Terms*—Constructivist learning theory, knowledge network, latent Dirichlet allocation (LDA), ontology relation learning, random walk.**

I. Introduction

Text symbol is the most important tool for human to express and share knowledge. However, it is still a very hard task on how to make machine autonomously understand the semantic information of concept symbol from human corpus [1], [2]. By borrowing the idea of ontology philosophy, ontology relation representation and learning methods using triples were introduced to model the semantic relations among concept symbols in previous research [3]. These relations mainly include hyponymy, synonymy, and antonymy relations. In order to learn these ontology relations, rule-based and statistical analysis methods are the two main strategies [4]–[6]. However, the above three types of ontology relations may not cover all requirements of ontology understanding, especially for concept demarcation. Besides, the latent Dirichlet allocation (LDA) [7], [8] model is also a famous tool to mine hidden semantic structures among knowledge concepts from text corpus. Even so, existing LDA models did not aim to mine explanation relations among knowledge concepts.

For the ontology relation learning problem, there are two other related techniques: 1) explanation-based learning [9] and 2) word embedding learning [10]. Explanation-based learning may be looked as a data-driven search space reduction. It assumes that real knowledge rules could be deduced based on priori domain axioms, all possible conclusions, and current observations. Wherein, ultimate knowledge conclusions should optimally explain all actual observations. Word embedding is a fundamental tool of using an embedded feature vector to represent the hidden semantic information of a knowledge concept. Then, the semantic relations might be uniformly computed in their embedded feature space.

In this article, a constructivist ontology relation learning (CORL) method is proposed inspired by the constructivist learning theory [11], [12]. In CORL, we consider that semantic generative process of documents (DSGP) is a progressive explanation-based associative process on a constructivist knowledge network. Wherein, each network node denotes a symbol concept, and its ontology is represented by a group of associative relations with other node concepts. Furthermore, each document could be modeled by a certain random walk process in DSGP, and constructivist ontology relations can be learned by optimally explaining given corpus based on the DSGP model.

Manuscript received September 23, 2019; revised May 30, 2020 and November 9, 2020; accepted December 19, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61872166; in part by the Six Talent Peaks Project of Jiangsu Province under Grant 2019 XYDXX-161; and in part by the Science and Technology Planning Project of Jiangsu Province under Grant BE2018056. This article was recommended by Associate Editor S. Ventura. *(Corresponding author: Zhenping Xie.)*

The authors are with the School of Artificial Intelligence and Computer Science and the Jiangsu Key Laboratory of Media Design and Software Technology, Jiangnan University, Wuxi 214122, Jiangsu, China (e-mail: xiezhenping@hotmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3138452.

Digital Object Identifier 10.1109/TCYB.2021.3138452

II. Related Works

A. Ontology Relation Learning

The concept of ontology [2] originates from philosophy research, which is used to denote the nature of object existence. In knowledge engineering research, Neches [13] first introduced and defined the ontology concept to depict a knowledge system. Knowledge ontology is also a basic description manner for the semantic nature and relation of lots of domain knowledge concepts. In general, an ontology system may contain a vocabulary and a group of logical statements to explain the semantic information and relations among knowledge concepts.

At present, semantic relation representation is commonly considered as two types, including: 1) taxonomic relations such as is-a (i.e., hypernym/hyponym) and 2) nontaxonomic relations. Nontaxonomic relations mainly are concept hierarchy relations, including inverse, transfer, classification, inheritance, instantiation, part-whole, attribute value, and so on.

B. LDA Model

The LDA model [7], [8] is a most famous document modeling tool, which can be used to topic word analysis [14], text classification [15], collaborative filtering [16], document retrieval [17] etc.

The LDA model is an unsupervised tool to model generative process of documents. It uses a three layers Bayesain probabilistic model to identify latent topic information of a group of document sets in corpus. By using the idea of word bag [18], the LDA model first regards each document as a word frequency vector. Then, a three-layer probabilistic model is considered as follows. Each document is represented by a probability distribution on some topics, and each

topic is represented by a probability distribution on a group of key words.

As an extension, hierarchical LDA models mainly aim at modeling superior or subordinate relations among key words, such as what hierarchical LDA [19] and latest L2H model [20] considered.

C. Constructivist Learning Theory

The constructivist learning theory comes from the idea of constructivism of explaining individual cognitive development process. It holds that human knowledge learning is a gradual constructing process, but not a simple data accumulation process. That is, new knowledge formation must depend on new abstraction and construction based on individual known knowledge and/or new observation data.

The constructivist learning theory emphasizes the essentiality of internal self-constitution when people learn to grasp new knowledge. Wherein, internal self-constitution consists of two aspects: 1) a knowledge ontology can be represented by other knowledge concepts in their symbol system and 2) knowledge growth must be an active progressive constructive process. For example, we can construct a new knowledge D by organizing existing knowledge A, B, and C and new observation data e. The above feature should be paid high attention for developing human-like machine knowledge learning systems.

At present, the constructivist learning theory is only a concept term mainly discussed in the education field, and has not been well concerned in machine knowledge learning.

D. Word Embedding Learning

In order to quantify semantic similarities between concepts, the word embedding representation has been well developed, including Word2 vec [21], ELMo [22], BERT [23], etc. These methods map any word into a high-dimensional vector by means of model training on certain corpus. Then, semantic relations between two words could be represented by computing their vector relations.

In this study, we attempt to directly use the concept word set itself to compose representation vector space. The semantic representation of any word is denoted by other words with constructive degrees. The constructive degree values reflect the associative explanation intensity from one word to another.

III. CORL Framework

A conceptual comparison between the learning problems of LDA and CORL was first given in Fig. 1. Wherein, the directions of arrows denote directed representation relations among knowledge concepts.

As shown in Fig. 1, the learning objective of LDA is to obtain the probabilistic relations between a document and its topics, and the probabilistic relations between every topic concept and its topic description words for a given corpus. Differently, the learning objective of CORL is to obtain the constructivist relations among all concept words. Wherein, we introduce associative probability to express the degrees of constructivist relations. Without doubt, concept words in CORL equally play two roles: 1) topic concepts and 2) topic description words, if similar explanation logic is used like in the standard LDA model. The above consideration may be looked as an extension of the relations considered in the LDA model. Furthermore, the total relation network structure, called the constructivist knowledge network (CKN), could be learned from a given corpus using LDA-like methods. Here, for convenience of explanation, we also use similar explanation logic like in current topic learning models. But in the CORL model, we think that ontology relation should be pure associative support relations among concept words (no specific topic or hierarchy for all concept words), which is first assumed in related research.

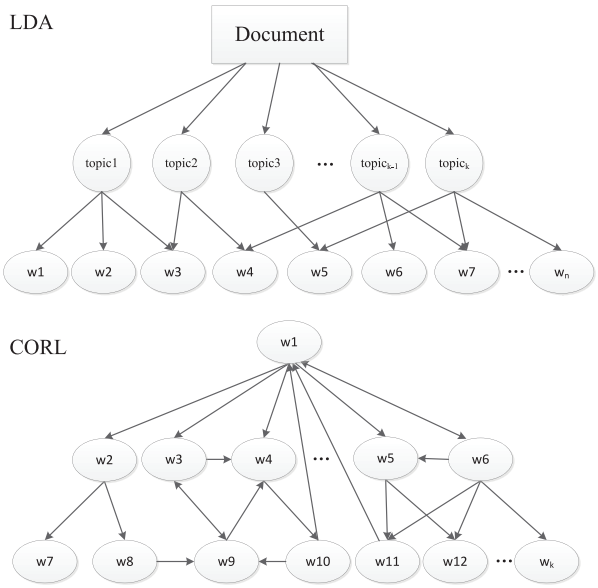

Fig. 1. Learning problem illustration of LDA and CORL.

In CKN, each node denotes a concept word, and edges among nodes denote constructivist relations among concept words. All node words pointed by a common source node are called the constructivist description words of that source node. For a directed link in CKN, a relation from a source node to a target node is called a forward associative relation, inversely a backward associative relation.

Furthermore, a DSGP is proposed in CORL using the random walk theory [24]. Wherein, the semantic process of a document could be represented by a group of ordered concept words. An ordered word sequence is looked as a generative process under certain random walk mechanism on document topics and CKN.

A. Associative Random Walk Mechanism

For any document, we may suppose that a group of ordered concept words is generated one by one based on author's associative thinking process based on initial document topic concepts. Such associative thinking process may be further regarded as an associative random walk mechanism (ARWM) on a domain CKN. In ARWM, if some nodes could be directly or indirectly connected in given CKN, then any ordered word sequence could be represented by a pathway in that CKN.

Fig. 2 gives a local illustration of ARWM, in which a concept word sequence $<\text{Nod}_1, \text{Nod}_2, \text{Nod}_3, \text{Nod}_4, \text{Nod}_5, \text{Nod}_6>$ is considered. Similar to Fig. 1, all thin gray directed arrows represent the forward associative relations. In addition, all thick red (solid) or blue (dashed) arrows represent, respectively, the direct or indirect jumps between two consecutive concept word nodes.

The above two types of jump ways between two consecutive concept words can be explained as follows.

1) *Directed Way:* Two consecutive words have a direct forward associative relation in CKN, like the jumps from Nod_1 to Nod_2, Nod_3 to Nod_4, and Nod_4 to Nod_5.
2) *Indirected Way:* Two consecutive words could establish a pathway by means of some intermediate nodes, like the pathways from Nod_2 to Nod_3, and Nod_5 to Nod_6.

Furthermore, we may suppose that for a given corpus, there should exist a connotative CKN that can optimally explain the generative

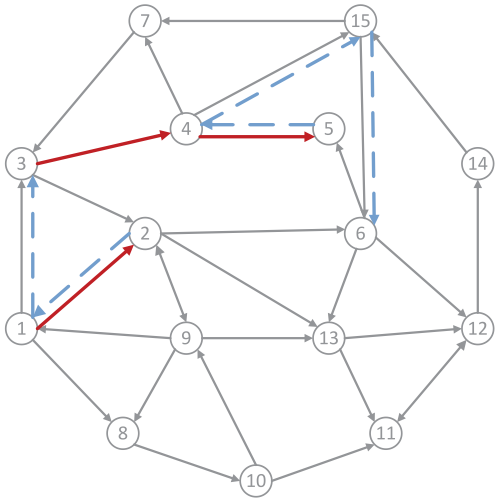

Fig. 2. Illustration of ARWM in CORL.

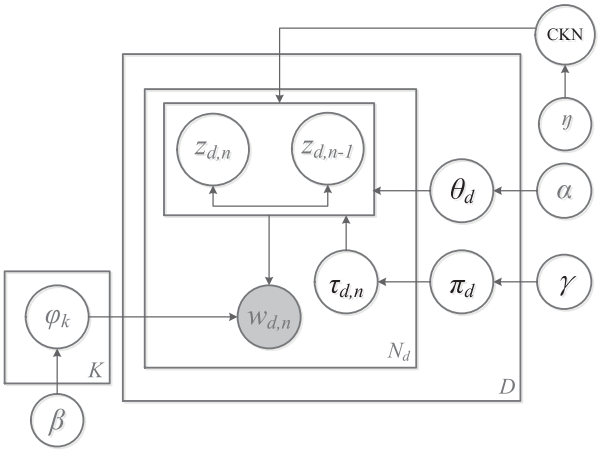

Fig. 3. ASGM.

process of all observed documents. Wherein, if the above generative model could be predefined, then corresponding CKN is also learnable.

B. Associative Semantic Generative Model

In order to extract meaningful knowledge structure from a set of domain documents (domain corpus), an associative semantic generative model (ASGM) was further introduced. ASGM mainly tries to use the sequence information among the observed words in given documents to extract the ontology relations of concept words.

In ASGM, a document is first considered to be composed of a group of prior topics like in the LDA model. Then, all observed document words are generated by corresponding topic concepts, and these topic concepts could be looked as topic concept sequences reflecting the associative thinking process of a document. Fig. 3 gives the framework description on proposed ASGM using a similar representation way of the classical LDA model.

In Fig. 3, K, D, and N_d, respectively, denote the numbers of all domain topics (also topic concepts according to our consideration), of all documents in given domain corpus, and of ordered concept words in an observed document d. β is the hyperparameter of the prior Dirichlet distribution on all domain topics. φ denotes the multinomial distribution on a group of topic description words of a same domain topic. That is, $\varphi \sim \mathrm{Dir}(\beta)$, the standard Dirichlet distribution. In CORL, φ also reflects the probabilistic distribution on the topic description words of a topic concept. It should be noticed that again, all concept words and all concept description words are a same word set.

Besides, θ_d represents the prior distribution on the document topics. α is also the hyperparameter of the prior Dirichlet distribution of θ_d. $z_{d,n}$ is a topic concept with respect to an observed document word $w_{d,n}$. η is the prior parameter of associative random walk in domain CKN. $\tau_{d,n}$ is a {0, 1} indicator variable to determine whether $z_{d,n}$ is generated depending on θ_d or $z_{d,n-1}$. π_d is the prior beta distribution of $\tau_{d,n}$ with hyperparameter $\gamma = (\gamma_0, \gamma_1)$.

Similar to the LDA model, the above variables are all latent expect that w is directly observable. In summary, the generative process of a document as well the observed document word sequence could be explained as follows.

1) For every document d, the model may generate a distribution on several prior topics with probabilistic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$.
2) For the nth ordered document word $w_{d,n}$, it is supposed to be generated from its topic concept $z_{d,n}$ with topic concept distribution $\varphi(z_{d,n})$.

According to the LDA model, $z_{d,n}$ is the topic concept of word $w_{d,n}$. Then, we may write

$$w_{d,n} \sim \mathrm{Mult}\big(\cdot|\varphi(z_{d,n})\big). \tag{1}$$

Wherein, Mult() is a multinomial distribution, and $z_{d,n}$ is the generative result according to document topic distribution θ_d if $\tau_{d,n} = 0$, or previous $z_{d,n-1}$ if $\tau_{d,n} = 1$. Thus, we may define

$$\begin{aligned} z_{d,n} &\sim p\big(z_{d,n}|z_{d,n-1}, \theta_d, \tau_{d,n}; \eta\big) \\ &= \begin{cases} p\big(z_{d,n}|\theta_d\big), & \tau_{d,n} = 0 \\ p_{\mathrm{hop}}\big(z_{d,n}|z_{d,n-1}; \eta\big), & \tau_{d,n} = 1 \end{cases} \end{aligned} \tag{2}$$

and

$$p\big(z_{d,n}|\theta_d\big) = \mathrm{Mult}(\theta_d) \tag{3}$$

$$p_{\mathrm{hop}}\big(z_{d,n}|z_{d,n-1}; \eta\big) = \frac{1}{J}\exp\left(-\frac{\mathrm{Hop}\big(z_{d,n}, z_{d,n-1}\big)}{\eta}\right). \tag{4}$$

Wherein, J is a normalization factor. $\mathrm{Hop}(z_{d,n}, z_{d,n-1})$ denotes the hops of the optimal walk pathway from $z_{d,n-1}$ to $z_{d,n}$ in CKN.

Compared to the classical LDA model, (4) is newly introduced, and (2) is modified. So, CORL may be equivalent to the classical LDA model in mathematics if we omit the influence of (4) that reflects the ordered sequence information of observed document words.

IV. Model Implementation

A. Main Procedure

In this section, we will discuss that how CORL can learn its hidden variables with given domain corpus. By analyzing the framework of the CORL method, two following procedures must be carried out.

1) *CKN Initialization:* This procedure is to constitute a basic connection structure with respect to a domain corpus. Wherein, domain topic concepts or topic description words should be first extracted. Then, initial connection relations could be preconstructed.
2) *Constructivist Ontology Relation Training:* This procedure is similar to the LDA model. All unknown hidden variables in CORL should be trained to optimally fit to observed corpus. At the same time, the distribution relations φ can be obtained.

B. CKN Initialization

1) Concept Word Extraction: This module is the first step of initializing CKN. For a given corpus, concept words could be extracted

by information entropy [25], word frequency [26], or other feasible methods. Before that, text denoising, word segmentation, and stop words removing should be preprocessed. In our experiments, information entropy and word frequency were used together, which is adequate for our experimental study according to our prior analysis. In theory, other methods also could be considered for different practical applications.

2) Connection Initialization: In order to produce reasonable initial relations among domain concept words, a word vector tool word2 vec [21] is used. For all extracted concept words, their 200-dimensional word vectors are first trained by means of the CBOW word2vec method, in which ten context words are used. Then, the distances between a word and other words are calculated using cosine distance.

Furthermore, we may choose a group of words for each word with top-k hight similarity to form initial connection relations. According to the research in small world networks [27] and our experimental analysis, the number of top-20 is set as default in this study.

Here, we only use the word2vec model (not complex ELMo or BERT) to create initial connection relations mainly because that the initialization quality using word2vec seems to be eligible. Moreover, our study mainly focuses on the subsequent constructivist relation learning among concept words. The core objective that we study in the CKN model is to exploit novel ontology relation representing model different from the current word embedding learning idea. In addition, ELMo and BERT are designed to fit the requirements of concrete complex context understanding applications, and such requirements are still not considered in CKN.

C. Model Training

Similar to the L2H method [20], Gibbs sampling algorithm [28] could be employed to inference CORL method. Given a domain corpus consisting of all observed document words $\{w_{d,n}\}$ and initial CKN structure, we might use the following interactive sampling procedure to estimate optimal CKN model parameters $\{\varphi_k\}$.

1) Sampling indicator variables $\{\tau_{d,n}\}$ using

$$p\big(\tau_{d,n}=j|w_d,z_{-d,n},\varphi\big)\propto\frac{C_{d,j}^{-d,n}+\gamma_j}{C_{d,\cdot}^{-d,n}+\gamma_0+\gamma_1} \tag{5}$$

except for $\tau_{d,0}\equiv 0$.

2) Sampling latent topic concept assignment $\{z_{d,n}\}$ according to the following probability formulas:

$$p\big(z_{d,n}|\tau_{d,n}=0,w_{d,n},z_{-d,n},\varphi\big)\propto\frac{N_{d,k|0}^{-d,n}+\alpha_k}{C_{d,0}^{-d,n}+\alpha_k}\times\varphi_{k,w_{d,n}} \tag{6}$$

$$p\big(z_{d,n}|\tau_{d,n}=1,w_{d,n},z_{-d,n},\varphi\big)\propto\frac{p_{\text{hop}}\big(z_{d,n},z_{d,n-1}\big)}{\sum_{i\in\{\leftarrow w_{d,n}\}}p_{\text{hop}}\big(z_i,z_{d,n-1}\big)}\times\varphi_{k,w_{d,n}}. \tag{7}$$

3) Estimating word distribution $\{\varphi_k\}$ using the following equation same as in standard LDA:

$$\varphi_{k,t}=\frac{N_k^t+\beta_t}{\sum_{t=1}^{K}N_k^t+\beta_t}. \tag{8}$$

Wherein, N_k^t denotes the frequency number of word t in topic k, $C_{d,j}$ denotes the total frequency number of words with indicator class j in the dth document from given text corpus, and $N_{d,k}$ represents the frequency number of all words in topic k. Then, the CORL's model training algorithm can be concluded in Algorithm 1.

Algorithm 1 CORL's Model Training Algorithm

```
Intialize CKN using gvien corpus S = {S_1, S_2, ..., S_D}
Set the total number iTera of iterative sampling traing, i = 0 and
  initialize {φ_k}_{k≤K}, γ_0 and γ_1
while i < iTera
  for d = 1 ... D
    for each observed word jump w_{d,n} → w_{d,n+1} in
      document d, do
      sample τ_{d,n} according to (5) and τ_{d,0} ≡ 0
      sample z_{d,n} according to (6) and (7)
      update the word distribution φ_k according to new z_{d,n} using
        equation (8)
    end
end
output top-k item for every topic k according to the weights
```

The above model training framework is similar to the training framework introduced in the L2H method. For CORL, more self-contained constructivist relations can be learned to obtain the effective ontology demarcation of each knowledge concept, which may be viewed as a further extension of L2H. Moreover, the distribution relations of a group of concept words to a concept word will be not only the statistical correlations but also the semantic explanative relations.

V. Experimental Studies

A. Performance Evaluation Method

Here, two kinds of ways are considered to evaluate model performance, including: 1) statistical performance evaluation and 2) case study [29]. Wherein, statistical performance evaluation methods are used to examine the reasonability of relation results obtained by CORL in the whole. Besides, the case study as auxiliary manner is used to visually show the validity of learning ability of CORL.

Concretely, statistical performance evaluation includes perplexity index [30], the accuracy of predicting document words, and indirect application performance.

Perplexity index is a measurement tool to reflect the degree that whether a sample cannot be predicted according to certain probabilistic distribution. For text modeling, the performance will be better if the perplexity value is smaller. For a given test document set $\tilde{S}=\{\tilde{S}_d\}$ and obtained φ, the perplexity value could be computed by

$$Pp(\tilde{S})=\exp\left\{\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}-\log p(w_{d,n})}{\sum_{d=1}^{M}N_d}\right\}. \tag{9}$$

Wherein, D is the number of test documents, and N_d is the number of observed words in document d. The computing formula of $p(w_{d,n})$ may be

$$p\big(w_{d,n}\big)=\int_{z,\tau}\varphi\big(w_{d,n}|z\big)p\Big(z|\tilde{S}_d,\tau\Big)dzd\tau \tag{10}$$

$$p\Big(z|\tilde{S}_d,\tau\Big)=\begin{cases}p\Big(z|\tilde{S}_d\Big), & \tau=0\\ p_{\text{hop}}\big(z,z_{d,n-1}\big), & \tau=1.\end{cases} \tag{11}$$

In (11), $p(z|\tilde{S}_d)$ is the document topic distribution same as in the standard LDA model. In this study, we choose 80% documents to train the model, and choose 20% documents to test model performance.

In addition, we also introduce an auxiliary evaluation index for CORL, the accuracy of predicting document words with a given

starting term concept word

$$A_{pr}\left(\tilde{S}_d\right) = \frac{1}{T}\sum_{i=1}^{T}\frac{\text{Count}\left(\left\{\widehat{w}_d^{\,i}\,|w_{d,0},\varphi\right\}\cap\{w_d\}\right)}{\text{Count}(\{w_d\})}. \quad (12)$$

Wherein, $\{\widehat{w}_d^{\,i}\,|w_{d,0},\varphi\}$ represents once document words simulatedly generated with the same number of $\{w_d\}$ using ARWM with distribution φ on trained CKN.

For indirect application performance, document classification, retrieval, and summarization may all be feasible. In this study, the document summarization is adopted.

B. Experimental Preparations

In our experiments, we will use two text datasets to investigate basic modeling performance of CORL by comparing the standard LDA model and L2H model.

On the one hand, we choose a public news article dataset provided by the Sogou laboratory [31]. On the other hand, we collect 14 600 technical articles with similar domains, including "health knowledge," "dietary nutrition," and "dietary misconceptions" from scientific knowledge Websites containing "China food science and technology"[1] and "39 health."[2] These articles contain about 20 million of words. The above two datasets may be, respectively, called the Sogou news dataset and our healthy dataset. Wherein, the Sogou news dataset is a public dataset that is well used in NLP experiments. In addition, an extra dataset on healthy knowledge was created to verify the performance stability of CORL.

According to CORL implementation, domain concept words and initial CKN should be pretrained. For two datasets, the number of concept words in CKN is set as 1000. By means of the method presented in Section IV-B, the 1000 words with higher importance are extracted. Some representative domain term concept words contain, milk, bean curd, egg, tomato, fruit, apple, papaya, bean, jujube, cherry, lemon, vegetables, honey, durian, ...

Next, the word vectors of these domain concept words are trained using the word2vec model. The basic connections among these word nodes could be constructed by means of the strategy presented in Section IV-B too. In Table I, a case of the connecting relations between the concept *Liver Cancer* and its top-20 related description words is illustrated from the Sogou news dataset. Here, we select the concept *Liver Cancer* as sample because that it may be more widely and easily understood by readers.

As shown in Table I, all related words of concept *Liver Cancer* are reasonable. The bigger distance may reflect higher statistical correlation. However, such connection representation may not be optimal to explain the concept scope of concept *Liver Cancer*, because some low correlated words may have high-similar semantics.

TABLE I
INITIAL CONNECTION RELATIONS OF CONCEPT WORD *Liver Cancer* FROM SOGOU NEWS DATASET

Concept description words	Word vector distances	Concept description words	Word vector distances
malignant tumor	*0.9411*	*canceration*	*0.6915*
HBV	*0.9355*	*disease*	*0.6544*
viruses	*0.9309*	*spread*	*0.5512*
medicine	*0.9203*	*cells*	*0.3547*
cancer	*0.9089*	*liver metastasis*	*0.2886*
AFP	*0.8852*	*lymph*	*0.2475*
cirrhosis	*0.8211*	*serum*	*0.1984*
hepatitis	*0.8003*	*surgery*	*0.1021*
treatment	*0.7532*	*infect*	*0.0897*
organ	*0.7404*	*pain*	*0.0518*

TABLE II
CONSTRUCTIVIST RELATIONS OF CONCEPT WORD *Liver Cancer* OBTAINED BY CORL ON SOGOU NEWS DATASET

Concept description words	Associative probability values	Concept description words	Associative probability values
cancer	0.102 ± 0.005	*treatment*	0.040 ± 0.001
liver	0.088 ± 0.007	*surgery*	0.036 ± 0.003
hepatitis	0.083 ± 0.008	*chemotherapy*	0.033 ± 0.003
hepatitis B	0.075 ± 0.009	*organ*	0.020 ± 0.001
cirrhosis	0.068 ± 0.005	*cells*	0.029 ± 0.005
malignant tumor	0.064 ± 0.001	*radiotherapy*	0.027 ± 0.003
liver metastasis	0.059 ± 0.003	*lymph*	0.016 ± 0.001
pain	0.054 ± 0.007	*HBV*	0.015 ± 0.002
disease	0.052 ± 0.008	*antiviral*	0.008 ± 0.0007
canceration	0.051 ± 0.002	*spread*	0.006 ± 0.0002

C. Basic Performance Results

Here, we, respectively, use CORL, LDA, and L2H to model the above two datasets. In experiments, the iterative number of model training is set as 1000, and hyperparameters are set as $\alpha = 0.1$ and $\beta = 0.001$ for all three methods and $\eta = 8.0$ for CORL by our experiential comparisons. First, Table II gives a case result obtained by CORL on Sogou news dataset for concept word *liver cancer*, in which associative probability values are listed.

By comparing Tables I and II, some interesting results could be found. First, some high important correlated words still have higher associative probability values, for example, cancer, malignant tumor, and cirrhosis. Second, some key semantic explanation words are allocated with higher relevancy ranks than direct word vector similarity. These words include hepatitis B, liver metastasis, and pain, and they obviously have higher associative probability values from human knowledge. In contrast, some common-sense words viruses and medicine are allocated as lower ranks, which may also be more reasonable. In addition, three important description words: 1) liver; 2) chemotherapy; and 3) radiotherapy, are extra found. These three words also should be important to demarcate the concept *liver cancer*.

Moreover, if a language model is well trained, then it could better explain given training dataset (higher posterior probability and lower perplexity index value). So, the perplexity index performance results on all corpus articles of two experimental datasets are reported in Table III for three compared methods. Wherein, the average results are calculated based on 20 time of runs for every dataset.

From Table III, the CORL method obtains better performance compared to two other methods. In addition, because our healthy dataset

[1]http://www.tech-food.com/
[2]http://www.39.net

TABLE III
PERPLEXITY INDEX PERFORMANCE OBTAINED BY THREE COMPARED METHODS ON TWO DATASETS

	LDA	L2H	CORL
Sogou News Dataset	20445±15	16009±14	14687±20
Our Healty Dataset	12489±35	9833±21	9846±18

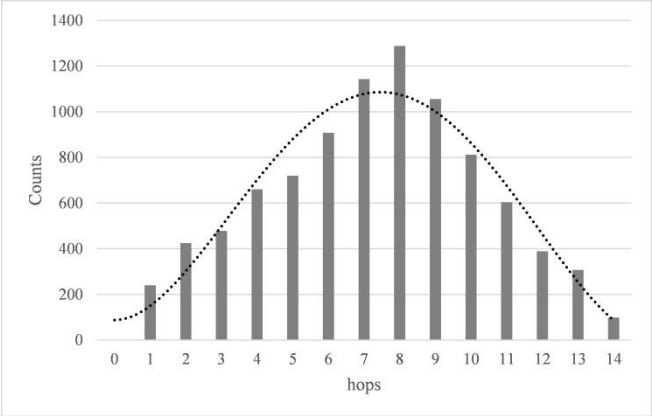

Fig. 4. Counts on different hops modeled by CORL on Sogou news dataset.

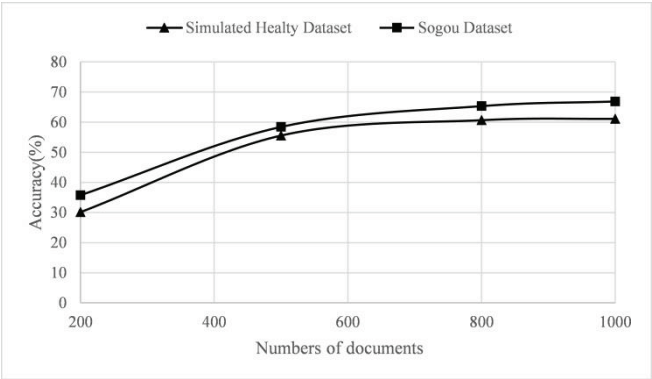

Fig. 5. A_{pr} performance obtained by CORL on two datasets.

is more subtly collected, the trained perplexity values are all lower than the public Sogou news dataset. The above results also reflect the modeling stability of CORL.

In Fig. 4, we further count the number of different random walk hops between two consecutive topic concepts on the Sogou news dataset. With the increase of hops, the counts will gradually rise till the maximum with $hops = 8$, and then decline steadily. Such statistical result might also reflect the reasonability of CORL method in a certain extent.

Furthermore, we analyze the accuracy of predicting document words for CORL on the above two datasets. The corresponding statistical results are shown in Fig. 5, in which two experimental datasets and different document numbers are considered.

From the results shown in Fig. 5, we may find that CORL could obtain relatively good prediction accuracy. The prediction accuracies can tend to stable values on two different datasets. Such results also reflect the effectiveness of our CORL method. For two different datasets, our healthy dataset presents slightly better performance than the Sogou dataset. The reason may be that our dataset only contains healthy domain's articles, but the Sogou dataset contains other domain articles.

TABLE IV
MOST RELATED DESCRIPTION WORDS OF CONCEPT *Liver Cancer* OBTAINED BY THREE COMPARED MODELS

LDA	L2H	CORL
cancer	*malignant tumor*	*cancer*
disease	*cancer*	*liver*
liver	*hepatitis B*	*hepatitis*
medicine	*organ*	*hepatitis B*
hepatitis B	*liver*	*cirrhosis*
viruses	*infected*	*malignant tumor*
hospital	*viruses*	*liver metastasis*
infected	*HBV*	*pain*
treatment	*treatment*	*disease*
organ	*medicine*	*canceration*

D. Concept Relation Analysis

Besides, if document models are trained on a given corpus, we may extract some most related words of a topic word to explain this topic. This objective is also the important semantic understanding requirement for NLP. For this purpose, a case result is exhibited in Table IV with respect to the term concept *liver cancer* on Sogou news dataset. For LDA and L2H methods, we extract top-10 related topic words of topic concept *liver cancer* by sorting their nearby degrees. For CORL, we extract top-10 related concept words from the concept word *liver cancer* to those words with top-10 high associative probability values in learned CKN.

By analyzing the results in Table IV, we may find that the extracted constructivist description words by CORL can better support the entire meanings of the concept *liver cancer*. That is, these words can compose a better associative word set. Specifically, the concept cirrhosis and pain should be very important constructivist explanations.

Moreover, we label the different words for three methods with underline in Table IV. Clearly, there are eight words that are same between LDA and L2H, while only four words are same between L2H and CORL. Concretely, the description words cirrhosis, liver metastasis, pain, and disease obtained by CORL seem to be more accurate explanation to concept *liver cancer* than words organ, viruses, treatment, and medicine obtained by L2H. In particular, the description words cirrhosis and pain are two very important concept descriptions for *liver cancer*. They are not preferentially found by LDA and L2H.

In addition, the description words have more clear hierarchies, for example, organ to liver, and viruses to HBV for the L2H method. Such results are consistent with the algorithm features of L2H. Different from LDA and CORL, some topic label information must be provided for L2H model training. However, LDA and CORL are more unsupervised.

At last, compared to LDA and L2H, CORL can obtain their description words for all concept words. But LDA and L2H only can be effective for a part of concept words. For LDA, only those topic words found by training procedure can obtain their description words. Similar to LDA, only those topic label words can obtain their description words for the L2H model. In this point, CORL is a more practicable method to learn overall ontology relations for a set of concept words with given corpus.

TABLE V
SOME TYPICAL ONTOLOGY RELATION SAMPLES OBTAINED BY L2H AND CORL

	L2H	CORL
inflation	null	*economics, currency, devaluation, price, price of commodities etc.*
skim milk	null	*lowfat milk, albumen, fat, milk, health etc.*
Yao Ming	null	*Houston Rockets, NBA, basketball, Houston, team etc.*
free kick	null	*football, foul, penalty shot, location, footballer etc.*
insurance	null	*society, guarantee, finance, insurer, premium etc.*
additives	*food additives*	*food additives, health, chemistry, nutrition, shelf life etc.*
lemon	*vitamin C*	*vitamin C, vitamin, antiphlogosis, whitening, orange etc.*

So, we further qualitatively examine ontology relation learning results obtained by L2H, LDA, and CORL. For the above Sogou news dataset, if same 1000 concept words are extracted as the total knowledge concept words, then for L2H only less half of concept words can be elected as valid topic concepts that they have their effective description words, and the number of topics obtained by LDA is much less. However, all concept words could be allocated to their effective description words for the CORL method. In Table V, some typical concept relation results obtained by L2H and CORL are exhibited.

As shown in Table V, for concept words inflation, skim milk, Yao Ming, free kick, and insurance, they did not be elected as topic concept words, and no effective description word was gained by L2H. Concretely, for the concept word inflation, it is a leaf node of its parent node word economics in gained concept hierarchy tree, and has no further subnode. Differently, the effective concept relation results also can be obtained by the CORL method for these concept words. Furthermore, for concept words additives and lemon, although few concept relation words of them could be extracted by L2H, but more abundant results could be obtained by the CORL method.

Based on the above results, new characteristics of CORL can be clearly demonstrated compared to existing related methods.

E. Application to Text Summarization

Text summarization [32] is an important tool to extract main contents from original documents. It is widely useful in text browsing, retrieval, and classification [33]. Here, we will, respectively, use CORL, LDA, and L2H methods to extract extra sentence features to better select abstract sentences.

Concretely, the maximum entropy classifier [34] in OpenNLP [35] is considered. The basic features of a sentence contain the sentence position, sentence length, and TF/IDF values of each key word in a sentence.

Moreover, we will consider the following strategy [33] to add sentence features. Generally speaking, if the topic concept of a sentence is more similar to the topics of its document, then this sentence will be more likely a summary sentence. Therefore, we can introduce the KL distance [36] between the topic distributions of a sentence and its document as added sentence feature. The corresponding computing formula could be defined as

$$\mathrm{KL}\big(p(z|s_{d,i})||p(z|S_d)\big) = \int_z p(z|s_{d,i}) \frac{p(z|S_d)}{p(z|s_{d,i})} dz. \tag{13}$$

Wherein, $p(z|s_{d,i})$ and $p(z|S_d)$, respectively, represent the topic distributions of a sentence $s_{d,i}$ and its document S_d. $s_{d,i}$ denotes the ith sentence in S_d. $p(z|S_d)$ could be directly obtained by CORL, LDA, and L2H models by training the models on entire document corpus $S_{1,2,\ldots,D}$.

TABLE VI
SUMMARY PERFORMANCE RESULTS OBTAINED BY FOUR COMPARED METHODS

	Pr	*Rr*
Baseline	55.98±0.15	41.34±0.11
LDA	75.36±0.25	70.58±0.34
L2H	83.16±0.11	77.89±0.24
CORL	83.74±0.23	76.44±0.12

Here, we choose a public document summary corpus provided by the Information Retrieval Research Laboratory of Harbin Institute of Technology [37]. In this dataset, there are a total of 1055 documents, and two types of summary sentences (with the ratio of 10% or 20%) are manually annotated. In the experiment, we randomly select 80% annotated summary sentences as training samples and other sentences as test samples. Moreover, 1000 topic concept words are preselected using the same strategy presented in Section IV-B. Similar to common text summarization research, two performance indices: 1) the precision Pr and 2) recall ratio Rr, are used.

Table VI lists the performance results obtained by four compared methods. The baseline results reflect basic performance without added KL distance feature. The other three rows of results denote the improved performance by three document modeling tools. Wherein, all performance values are statistical results of 20 time of runs.

From Table VI, we may know that the performance of the baseline method can be obviously improved if document modeling tools are added. In particular, the precision and recall ratio improved by the CORL method are similar to the performance obtained by the L2H method. This result can well indicate the effectiveness of CORL method.

VI. CONCLUSION

In this study, a novel ontology relation learning method, CORL, is proposed inspired by the constructivist learning theory. Wherein, associative probability values between concept words are used to express the constructivist intensity. In CORL, a novel ASGM is proposed to model document semantics.

In CORL, the model inference algorithm could be derived using a similar principle of the L2H method. In our experiments, some qualitative analysis and quantitative comparisons are performed. Corresponding results are noteworthy, and the effectiveness of CORL can be clearly verified.

In principle, CORL combines the ideas of knowledge graph and topic model, and can be viewed as a brand-new development of existing ontology learning methods. It expands the scope of ontology relation learning, and may be widely developed and used in more NLP problems except discussed problems in this article.

Nevertheless, the associative random walking mechanism in CORL should be further optimized, which is also our next important research work. In addition, for the problem how to effectively evaluate the performance of CORL, it is a trouble. According to the suggestions of

anonymous reviewer, maybe, it is possible to compare the associated words with a medical dictionary or thesaurus. However, this work may be very complicated because of the complexity of explanative text of medical concepts in a medical dictionary, and it could be specially studied in future work.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[2] W. Wong, W. Liu, and M. Bennamoun, "Ontology learning from text: A look back and into the future," *ACM Comput. Surveys*, vol. 44, no. 4, pp. 1–36, 2012.

[3] C. Sun, M. Zhao, and Y. Long, "Learning concepts and taxonomic relations by metric learning for regression," *Commun. Stat.*, vol. 43, no. 14, pp. 2938–2950, 2014.

[4] J. Punuru and J. Chen, "Learning non-taxonomical semantic relations from domain texts," *J. Intell. Inf. Syst.*, vol. 38, no. 1, pp. 191–207, 2012.

[5] I. Qasim, J.-W. Jeong, J.-U. Heu, and D.-H. Lee, "Concept map construction from text documents using affinity propagation," *J. Inf. Sci.*, vol. 39, no. 6, pp. 719–736, 2013.

[6] B. E. Idrissi, S. Baïna, and K. Baïna, "Automatic generation of ontology from data models: A practical evaluation of existing approaches," in *Proc. IEEE 7th Int. Conf. Res. Challenges Inf. Sci.*, 2013, pp. 241–252.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[8] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "A deep and autoregressive approach for topic modeling of multimodal data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1056–1069 Jun. 2016.

[9] T. G. Dietterich and N. S. Flann, "Explanation-based learning and reinforcement learning: A unified view," *Mach. Learn.*, vol. 28, no. 2, pp. 169–210, 1997.

[10] S. He, K. Liu, G. Ji, and J. Zhao, "Learning to represent knowledge graphs with Gaussian embedding," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 623–632.

[11] E. Murphy, *Constructivist Learning Theory*. Boston, MA, USA: Springer, 2012, p. 787.

[12] J. W. Zhang and Q. Chen, "From cognitive to constructivism," *J. Beijing Normal Univ. Social Sci.*, vol. 136, no. 4, pp. 75–82, Jul. 1996.

[13] R. Neches *et al.*, "Enabling technology for knowledge sharing," *AI Mag.*, 1991, vol. 12, no. 3, pp. 36–56, 1991.

[14] H. Çelikkanat, G. Orhan, N. Pugeault, F. Guerin, E. Sahin, and S. Kalkan "Learning context on a humanoid robot using incremental latent Dirichlet allocation," *IEEE Trans. Cogn. Devel. Syst.*, vol. 8, no. 1, pp. 42–59, Mar. 2016.

[15] M. Pavlinek and V. Podgorelec, *Text Classification Method Based on Self-Training and LDA Topic Models*. Oxford, U.K.: Pergamon Press, Inc., 2017.

[16] X. Zhou and S. Wu, *Rating LDA Model for Collaborative Filtering*. Amsterdam, The Netherlands: Elsevier Sci. Publ. B.V., 2016.

[17] D. Ganguly, J. Leveling, and G. J. F. Jones, "An LDA-smoothed relevance model for document expansion: A case study for spoken document retrieval," in *Proc. ACM SIGIR*, 2013, pp. 1057–1060.

[18] Y. Ko, "A study of term weighting schemes using class information for text classification," in *Proc. ACM SIGIR*, 2012, pp. 1029–1030.

[19] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 17–24, 2010.

[20] V.-A. Nguyen, J. L. Boyd-Graber, P. Resnik, and J. Chang, "Learning a concept hierarchy from multi-labeled documents," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2014, pp. 3671–3679.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR Workshop*, 2013, pp. 1–12.

[22] M. E. Peters *et al.*, "Deep contextualized word representations," in *Proc. NAACL-HLT*, 2018, pp. 2227–2237.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. AACL-HLT*, vol. 1, 2019, pp. 4171–4186.

[24] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.

[25] J. A. Núñez, P. M. Cincotta, and F. C. Wachlin, "Information entropy," *Celestial Mech. Dyn. Astron.*, vol. 64, nos. 1–2, pp. 43–53, 1996.

[26] R. H. Baayen and R. Lieber, "Word frequency distributions and lexical semantics," *Comput. Humanities*, vol. 30, no. 4, pp. 281–291, 1996.

[27] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.

[28] A E. Gelfand, "Gibbs Sampling," *J. Amer. Stat. Assoc.*, vol. 95, no. 452, pp. 1300–1304, 2000.

[29] M. S. Antoniou, "Case study research: Design and methods," *Eval. Res. Educ.*, vol. 24, no. 3, pp. 221–222, 2003.

[30] J. Horgan, "From complexity to perplexity," *Sci. Amer.*, vol. 272, no. 6, pp. 104–109, 1995.

[31] "Sogou Labs [EB/OL]." [Online]. Available: https://www.sogou.com/labs/ (Accessed: Jun. 20, 2018).

[32] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proc. ACM SIGIR*, 1995, pp. 68–73.

[33] A. Nenkova and K. Mckeown, "A survey of text summarization techniques," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 43–76.

[34] "Maximum Entropy Modeling [EB/OL]." [Online]. Available: http://homepages.inf.ed.ac.uk/lzhang10/maxent.html (Accessed: Jun. 20, 2018).

[35] "Apache OpenNLP [EB/OL]." [Online]. Available: http://opennlp.apache.org/ (Accessed: Jun. 20, 2018).

[36] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Process.*, vol. 11, pp. 146–158, 2002.

[37] "Information Retrieval Research Laboratory of Harbin Institute of Technology [EB/OL]." [Online]. Available: http://ir.hit.edu.cn/ (Accessed: Jun. 20, 2018).