Can Machines Intelligently Propose Novel and Reasonable Scientific Hypotheses?

Pengwei Wang
South China University of Technology
w.pengwei@mail.scut.edu.cn

Zhongyuan Wang
Facebook Inc.
zhy@fb.com

Lei Ji
Microsoft Research
leiji@microsoft.com

Jun Yan
Microsoft Research
junyan@microsoft.com

Lianwen Jin
South China University of Technology
lianwen.jin@gmail.com

ABSTRACT

Machine intelligence is attracting increasing attention from both industry and academia. However, the problem of how to make machines innovate novel hypothesis is underexplored. Automatic hypothesis generation can effectively shorten research process. In this work, we try to build an embedding based genetic algorithm to learn "experience" from past data, mine latent semantic information, and then propose the new scientific hypotheses. To our best knowledge, we are the first who propose to use an embedding based genetic algorithm for scientific hypothesis generation. Experiments show that our method outperforms the state of the art.

1. INTRODUCTION

Many people would like to see into the future. However, it is hard to process all known knowledge to discover unknown knowledge for human beings. Validating whole trials needs much time and money. In this paper, we are interested in how to let machines propose novel and reasonable ***scientific hypotheses***, which can efficiently accelerate scientific progress. Based on Wikipedia, scientific hypothesis is referred to a trial solution (***S***) to a problem (***P***).

In brief, a hypothesis is defined as a triple (***P***,***R***,***S***) in this paper. ***R*** denotes a relation between ***P*** and ***S***. Our goal is to extract the known triples from past literature, and outputs new hypotheses which have high probability to be validated by experts in the future. However, previous work [4] misses the latent semantic relationships among solutions or problems, and cannot predict hypotheses containing entities out of existing knowledge graph. Intuitively, the generation of new hypotheses can be simulated as a process of natural evolution. Inspired by this, we think to use an evolutional algorithm, ***genetic algorithm*** (*GA*) which evolves toward better individuals, to finish this task. The chromosomes of individuals (triples) are firstly learned. We propose a ***joint text and knowledge graph embedding model*** to encode semantic information into chromosomes. Then, the evolutional process is iterated until the algorithm converges and some highly fit individuals (hypotheses) are survived. Finally, we recover the names of the new individuals. Our experiments show that our method outperforms existing work.

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054244

2. TASK AND NOTATION DEFINITION

We let knowledge graph **D** stand for a set of triples in the form (p, r, s), $p \in \mathbf{P}$, $s \in \mathbf{S}$, $r \in \mathbf{R}$, where **P** is the problem entity set, **S** is the solution entity set and **R** is a relation set. We use bold letters **p**, **r**, **s** to denote the embedding of p, r, s. We denote $\mathbf{E} = \mathbf{P} \cup \mathbf{S}$ as the whole knowledge graph entity set. **W** is denoted as the words set in domain knowledge related free text corpus.

In detail, as shown in Figure 1, our task takes $\{[\mathbf{p}, \mathbf{r}, \mathbf{s}] | (p, r, s) \in \mathbf{D}\}$ as input, and outputs new triple embedding $\{[\mathbf{p}^{'}, \mathbf{r}^{'}, \mathbf{s}^{'}]\}$. In particular, since there is no new relation generation in this work, we keep r fixed. Then $\mathbf{E}$ and $\mathbf{W}$ are used to search the name of $\mathbf{p}^{'}, \mathbf{s}^{'}$. Finally, new hypotheses $(\boldsymbol{P}, \boldsymbol{R}, \boldsymbol{S})$ are generated, where $\boldsymbol{P}$, $\boldsymbol{R}$ and $\boldsymbol{S}$ are the corresponding names of $\mathbf{p}^{'}$, $\mathbf{r}^{'}$ and $\mathbf{s}^{'}$.

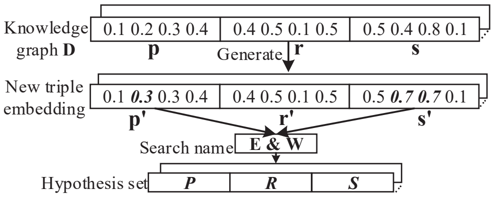


Figure 1: The flowchart of the task .

3. HYPOTHESIS GENERATION

In our framework, chromosomes of individuals are learned at first. And then, to simulate the natural evolution, new chromosomes are generated using GA. Finally, we search the name of the generated chromosomes in domain knowledge related free text.

3.1 Chromosomes Learning

We propose three methods to learn the chromosomes considering the knowledge graph and domain knowledge related free text.

KGEGA: The chromosome of a individual (p, r, s) can be represented as $[\mathbf{p}, \mathbf{r}, \mathbf{s}]$. Inspired by this, we propose to use knowledge graph embedding [1] to generate **p**, **r** and **s**, which is named as ***knowledge graph embedding based GA model*** (***KGEGA***).

TEGA: Considering the entity coverage of KGEGA, we propose to learn the chromosomes from free text using text embedding [3], which is named as ***text embedding based GA model*** (***TEGA***).

JEGA: To keep both the precision and coverage, we propose to encode the text into the knowledge graph embedding. We use a one layer neural network applied to the text embedding of **W**. The output of the neural network is used to construct the knowledge graph embedding using equation (1), which enforces that the **s** should be close to the **p** plus the **r** if (p, r, s) holds:

$$S_{kb}(p, r, s) = \| \mathbf{p} + \mathbf{r} - \mathbf{s} \|_{l_1/l_2} \quad (1)$$

$$\mathbf{p} = \mathrm{sigmoid}(W\mathbf{p}_{_T} + \mathbf{b}) \quad \mathbf{s} = \mathrm{sigmoid}(W\mathbf{s}_{_T} + \mathbf{b}) \quad (2)$$

where $\mathbf{p}_{_T}$ and $\mathbf{s}_{_T}$ are the vector representation of p and s in text embedding space. The W and $\mathbf{b}$ are the parameters. We name this model as ***joint embedding based GA model*** (***JEGA***).

3.2 New Chromosomes Generation

This part takes the chromosomes as input, and outputs new chromosomes using GA. In each iteration, the new population is generated by repeating selection, crossover and mutation until the number of the new population and current population is same. We select $\| \mathbf{p} + \mathbf{r} - \mathbf{s} \|_{l_2}$ as fitness function. Finally, new highly fit chromosomes will be generated after the algorithm converges.

3.3 Hypothesis Generation

This part takes the new chromosomes, knowledge graph **D** and domain knowledge related free text as input, and outputs the hypotheses. We take the first and last third of the new chromosome as a problem vector and a solution vector. The entity (***P*** or ***S***) name can be found from the **D** if the corresponding vector exists in the knowledge graph embedding. Otherwise, we propose to use the nearest neighbor to search the names in text by evaluating the similarity between the entity vector and the text embedding of **W**.

4. EXPERIMENTS

4.1 Data Collection

In this work, we use two datasets: DBLP and PubMed. We take the papers published before 2014 as training data, and the papers published from 2014 to now as evaluation data. Then, the evaluation data are randomly split into test data and validation data halves. (***P***,***R***,***S***) denote ("problem","solved-by","algorithm") and ("disease","treated-by","drug") in DBLP and PubMed. (***P***,***R***,***S***) is simply regarded as a triple when ***P*** and ***S*** co-occur in a sentence. Entities are tagged by a NER tool[1]. About the domain knowledge related free text, we obtain web sentences containing entities in DBLP and PubMed from a commercial search engine. We also crawl question answering data from two websites[2].

4.2 Results

The dimension size of knowledge graph and text embedding is set as 300. The cross and mutation rate in GA are 0.001 and 0.005 respectively. The initial population size is set as the size of **D**.

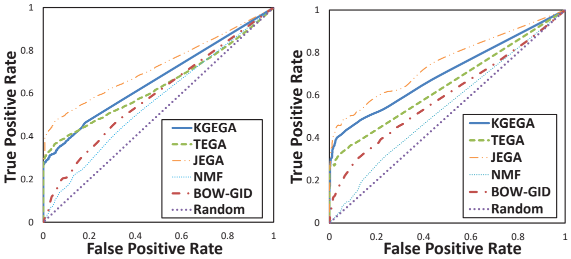
Link prediction: This task is designed to complete a triple (p,r,s) with p or s missing. We use *Mean Rank* and *Hits@10* as metric to evaluate the performance of the learned chromosomes.

We compare **JEGA** with other models. The results are shown in Table 1. The JEGA obtains the best result. Our method can be easily applied to new knowledge graph embedding models.

Table 1: Link prediction results.

Datasets	DBLP		PubMed	
Metric	Mean Rank	Hit@10(%)	Mean Rank	Hit@10(%)
RESCAL[5]	789	14.70	1045	7.41
LFM[2]	341	36.58	423	36.78
Skip-gram[3]	351	39.45	401	45.37
TransE[1]	242	51.46	321	47.89
JEGA	**124**	**59.63**	**261**	**58.16**

Compare with previous hypothesis generation models: Compared with NMF [4] and BOW-GID [4], the results are shown in Figure 2. We use Receiver Operating Characteristic curves to measure performance. From the results, we can see that our proposed methods (KGEGA, TEGA and JEGA) outperform other models, and the JEGA achieves the best result. This also indicates that our methods can efficiently generate hypotheses which are very likely to be verified in the future.



(a) DBLP result. (b) PubMed result.

Figure 2: Comparison with previous work

Case Studies: Finally, we show other new hypotheses in DBLP and PubMed with high confidence predicted by our model in Table 2 and Table 3 respectively, which are still not yet verified.

Table 2: New predicted hypotheses in DBLP

Problem	Relation	Trial Solution
fault prediction	*Solved-by*	*sparse coding*
dialogue systems	*Solved-by*	*manifold embedding*
knowledge graph embedding	*Solved-by*	*tensor decomposition*
text summarization	*Solved-by*	*sparse coding neural network*
video compression	*Solved-by*	*huffman coding*

Table 3: New predicted hypotheses in PubMed

Problem	Relation	Trial Solution
fibrosis	*Treated-by*	*oxygen*
pneumoperitoneum	*Treated-by*	*somatostatin*
lung cancer	*Treated-by*	*rituximab*
colon cancer	*Treated-by*	*neostigmine*
ovarian cancer	*Treated-by*	*streptokinase*

5. CONCLUSION AND FUTURE WORK

In conclusion, we present a framework for generating hypotheses using embedding based genetic algorithm. In future work, we will apply our methods to more domains. Besides, we will design more fitness functions. Finally, only one-to-one hypotheses are generated by our models in current work. We will combine multi-solutions to solve one problem in the future work.

6. ACKNOWLEDGMENTS

This research is supported in part by NSFC (Grant No.: 61472144), the National Key Research & Development Plan of China (Grant No.: 2016YFB1001405), GDSTP (Grant No.: 2015B010101004, 2015B010130003, 2015B010131004), GZSTP(Grant No.: 201607010227).

7. REFERENCES

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multirelational data. *NIPS*, 2013.

[2] R. Jenatton, N. L. Roux, A. Bordes, and G. Obozinski. A latent factor model for highly multi-relational data. *NIPS*, 2012.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*, 2013.

[4] M. Nagarajan et al. Predicting future scientific discoveries based on a networked analysis of the past literature. *SIGKDD*, pages 2019–2028, 2015.

[5] M. Nickel, V. Tresp, and H.-P. Kriegel. Factorizing yago: scalable machine learning for linked data. *WWW*, 2012.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

[2] http://stackoverflow.com/ http://www.drugs.com/