An Unsupervised Approach of Truth Discovery From Multi-Sourced Text Data

Chen Chang*, Jianjun Cao†, Guojun Lv* and Nianfeng Weng†
*Army Engineering University
Nanjing, China 210001
Email: c308051252@163.com, l1983286838@outlook.com
†The Sixty-third Research Institute, National University of Defense Technology
Email: jianjuncao@yeah.net, wengnf@hotmail.com

***Abstract*—In the era of information explosion, multi-sourced data may exist conflicts or even errors. To address this issue, plenty of truth discovery methods have been proposed to get trustworthy information from conflicting data. However, most existing truth discovery methods are designed for structured data and cannot meet the strong need to extract trustworthy information from raw text data. For text data, there are no completely correct or wrong answers, most answers may be partially correct. It is quite different from the situation of traditional truth discovery. In addition, tradition methods estimate the reliability of source based on plenty of observations provided. Unfortunately, for the scene of text truth discovery, it is not easy to get enough observations for sources. Besides, traditional methods ignore the importance of structure information and semantic information of text data, which leads to suboptimal results. To solve these challenges, we propose a Graph Convolutional Network (GCN) based truth discovery model to discover trustworthy information from text data. Firstly, Smooth Inverse Frequency (SIF) is utilized to learn real-valued vector representations for text data. Then, we construct undirected graph with these vectors to capture the structure information of answers. After that, the GCN is used to store and update the reliability of these answers, which sums up all the feature vectors of all neighboring answers to improve the accuracy and efficiency of truth discovery. Different from traditional methods, we use vectors to store the reliability of answers which have higher representation capability compared with real numbers, and network is adopted to capture complex relationships between answers rather than simplified functions. The experiment results on real datasets show that though text data structures are complex, our model can still find reliable answers compared with retrieval-based and state-of-the-art approaches.**

I. Introduction

The era of big data has arrived, the amount of information on the Internet has grown rapidly, and tremendous data can be provided by different online platforms (i.e., Yahoo Answers). Due to the openness and multi-source of the Internet, such information may exit errors or even conflicts. Therefore, how to get reliable information from low-quality multi-sourced data is a challenging problem.

Various of methods proposed to estimate the reliability of source and infer trustworthy information form noisy generated data. However, most existing truth discovery methods are designed for structured data, and can not meet the strong need to extract trustworthy information from text data which has its unique natural language characteristics. First, the answers to a question may be multifactorial, and it is usually hard for a given answer to cover all the factors. For a question like "What are the symptoms of flu?", the answer contains fever, chills, cough, nasal symptom, etc. If a user provides two factors of the correct factors, the existing truth discovery methods may determine this answer to be completely wrong and assign a low reliability to this user. Second, answers provided by online users may convey a very similar meaning with different expressions. For example, users may use words such as tired or exhausted to describe the symptom of fatigue, but existing truth discovery methods may treat them as totally different two words.

In addition, there are some flows for traditional truth discovery methods that constrain the results of truth discovery. First, traditional truth discovery methods estimate source reliability based on plenty of observations source provided. For the text data truth discovery, same problem may be answered by many users, while the users usually answer few questions. The sparseness of data may cause difficulty in user reliability estimation. Traditional methods relying on users' numerous observations for reliability estimation is not applicable to text data truth discovery. Second, previous works in truth discovery make a common assumption that the source-claim relational dependency can be represented by some simplified functions (i.e., linear, quadratic, binomial). This assumption can easily lead to suboptimal truth discovery results because neither the answer credibility nor the source reliability is known a priori. Third, traditional methods fail to fully use the structure information from answers but discover truth separately. In other words, tradition methods fail to fuse information from large number of observations.

In this paper, we propose a GCN based model that fits for the challenges to infer trustworthy information from text data. The major contributions of this paper are:

- To the best of our knowledge, we are among the first to construct graphs and develop neural network for solving text truth discovery problem.
- The proposed model can accurately capture the answers reliability without casting it to some over-simplified functions compared with traditional truth discovery methods.
- The structure information of all answers is treated as latent background knowledge, each answer sums up all

978-1-7281-4528-0/19/$31.00 ©2019 IEEE
DOI 10.1109/DSC.2019.00046

the semantic information of all neighboring answers, and can be stored to help find credible answers.

II. Related Work

1) Truth Discovery from Structured Data: Various of methods proposed based on the principle that the sources providing true information more often will be assigned higher reliability degrees, and the information that is supported by reliable sources will be regarded as truths [1]. These truth discovery approaches can be roughly divided into the following four categories: iterative methods [2]–[4], in which the truth computation step and source reliability degree estimation step are iteratively conducted until convergence; optimization based methods [5]–[8], in which a distance function will be defined to measure the difference between the information provided by source and the identified truth; probabilistic graphical model based methods [9]–[11], in which each claimed value is generated based on the two parameters corresponding truth and source weight, expectation maximization is widely used to infer the latent variables; neural network based methods [12]–[15], in which neural network is applied to accurately estimate the source-claim relational dependency function, which may be very complex.

2) Truth Discovery from Text Data: For truth discovery from text data, most researchers have simplified the problem and only judge whether the text data is true. Popat et al. constructed the "source-language-style" vector as input vector, and converted the truth discovery problem into 0/1 classification problem. Logistic regression was used to deal with this classification problem [16]. Broelemann et al. used the hidden layer of restricted boltzmann machine to learn the probability distribution of latent truth. Due to the characteristic of the restricted boltzmann machine, this method can only deal with the data with binary attribute [17], Marshall et al. used the fully connected neural network to learn the relationship between the reliability of the source and the credibility of claims. This method also simplified the users' answers to be true or false [18]. To the best of our knowledge, only Li et al. [19] and Zhang et al. [20] make full use of the semantic information of text data and proposed the real text truth discovery methods. First method can only handle single word answers. Second method combines the keywords extracted from the answers of specific question into multiple interpretable factors, and used the method based on probabilistic graphical model to perform truth discovery to find trustworthy answers. However, the method made the assumptions about the distribution of observations, but in fact the relationship between answers is unknown and complex, an inappropriate distribution hypothesis will lead to an unsatisfied truth discovery results.

III. Framework

A. Problem Statement

In this paper, we consider a general truth discovery problem for text data. Before introducing the model, we first define the problem as follows.

Given a question q, a set of answers $A = \{a_i | i = 1, 2, \ldots, M\}$, where M denotes the number of answers for this question. These answers can be complex and partially correct. The purpose of this paper is to find highly-trustworthy answers for each question.

B. Vector Representation Learning

When applying truth discovery problem to text data, making full use of the semantic information of natural language is of vital importance. For this reason, we learn real-valued vector representations $x_i (i = 1, 2, ..., M)$ for answers $a_i (i = 1, 2, ...M)$ by SIF [29] to vectorize semantic information from answers. $X \in \mathbb{R}^{M \times K}$ stands for a matrix of all answer vectors for this question, where K denotes the dimension of answer vectors. We think that the relationship and semantic information between answers can be learned on the basis of the sentence-embedding-learning algorithm.

SIF is a simple approach for sentence embedding based on the discourse vectors in the random walk model for generating text [30]. It is simple and unsupervised, but achieves significantly better performance than baselines on various textual tasks, and can even beat sophisticated supervised methods such as some RNN and LSTM models.

C. Truth Discovery

In our model, we encode the graph structure and the semantic information of answers using GCN model $f(\cdot)$ and perform truth discovery on an unsupervised target. Conditioning $f(\cdot)$ of the graph will allow the model to distribute gradient information from unsupervised loss designed for text data truth discovery and learn representations of identified truth answers based on the assumption of truth discovery.

1) Construct Undirected Graph: In this step, We construct undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with M nodes $v_i \in \mathcal{V}$, eages $(v_i, v_j) \in \mathcal{E}$, an adjacency matrix $A \in \mathbb{R}^{M \times M}$ (binary), and a degree matrix $D_{ii} = \sum_j A_{ij}$. The answers are set as nodes of graph. Based on the assumption that connected nodes share the similar semantic information, when the similarity $s(x_i, x_j)$ between two answer vectors x_i and x_j is greater than the threshold $\alpha\, (0 \leq \alpha \leq 1)$, the two nodes are connected, $s(x_i, x_j)$ is defined as the normalized cosine similarity between two answer vectors:

$$s(x_i, x_j) = 1 - \frac{1}{\pi} \cos^{-1} \left(\frac{x_i \cdot x_j}{\sqrt{x_i} \sqrt{x_j}} \right), \tag{1}$$

where $s(x_i, x_j) \in [0, 1]$, when $s(x_i, x_j) = 0$ means two answers a_i and a_j are complete different, and $s(x_i, x_j) = 1$ means two answers a_i and a_j are most similar.

In fact, not all answers essential for the truth discovery process later. Observing that the correct answers usually have similar semantic information, the degree of such nodes are large, and the wrong answers are always different from each other, the degree of such nodes in the graph are usually small. According to this principle, We construct the subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ from $\mathcal{G}$ by setting the threshold $\beta (\beta > 0)$ of the degree. IF only the degree of the node v_i is larger than

β, the node is retained. Therefore, we compress the scale of the problem, save the running time of truth discovery, and effectively improve the convergence speed and accuracy of the truth discovery. $X' \in \mathbb{R}^{M' \times K}$ stands for the matrix of reserved answer vectors for this question, M' denotes the number of reserved answers, $A' \in \mathbb{R}^{M' \times M'}$ and $D_{ii} = \sum_j A'_{ij}$ stands for the adjacency matrix and degree matrix of graph $\mathcal{G}'$.

2) Truth discovery by GCN: Inspired from [31], we adapt flexible model $f(X', A')$ for efficient text truth discovery by the answer vector matrix X' and adjacency matrix A' of the graph $\mathcal{G}'$. By this way, it will be powerful in truth discovery where the adjacency matrix A' contains information of answers relations not present in the X'. The overall model of multi-layer GCN for unsupervised text data truth discovery is depicted in Fig. 1.

For our model, we consider a two-layer GCN for unsupervised discovering trustworthy answers and consider the following simple form of layer-wise propagation rule:

$$\begin{aligned} Z &= f(X', A') \\ &= \sigma\left(A'\sigma\left(A'X'W^{(0)}\right)W^{(1)}\right), \end{aligned} \tag{2}$$

where $W^{(0)}$ and $W^{(1)}$ are weight matrices for the two neural network layers, which can be considered to store reliability information of answers that is useful in truth discovery. In traditional methods, the reliability of answers is represented by a real number and is treated as a weight in computing the credibility of information. We vectorize the reliability of answers and treat them as weight matrices in evaluating the credibility of answers. $\sigma(\cdot)$ is a non-linear activation function (i.e., ReLU). Z denotes the output matrix on the basis of the semantic information X', reliability information $W(0), W(1)$, and structure information A'.

Although this model is already quite powerful, there are two limitations needs to be addressed. First, multiplication with A' means that, for every node, we sum up all the feature vectors of all neighboring nodes but not the node itself (unless there are self-loops in the graph). We can "fix" this by enforcing self-loops in the graph: we simply add the identity matrix to A'. Second, A' is typically not normalized and therefore the multiplication with A' will completely change the scale of the feature vectors. Normalizing A' such that all rows sum to one gets rid of this problem. We use a symmetric normalization $D'^{-\frac{1}{2}}A'D'^{-\frac{1}{2}}$ to solve this problem.

Combining these two tricks, we essentially arrive at the propagation rule introduced as follows.

$$\begin{aligned} Z &= f(X', A') \\ &= \sigma(\tilde{D'}^{-\frac{1}{2}}\tilde{A'}\tilde{D'}^{-\frac{1}{2}}\sigma\left(\tilde{D'}^{-\frac{1}{2}}\tilde{A'}\tilde{D'}^{-\frac{1}{2}}X'W^{(0)}\right)W^{(1)}), \end{aligned} \tag{3}$$

where $\tilde{A'} = A' + I_N$, $W^{(0)} \in \mathbb{R}^{K \times H}$ is an input-to-hidden weight matrix for a hidden layer with H nodes, $W^{(1)} \in \mathbb{R}^{H \times K}$ is a hidden-to-output layers. Multiplication with $\tilde{A'}$ means that, for every node, we sum up all the semantic information of all neighboring nodes with the node itself.

For truth discovery problem, according to the hypothesis: 1) the reliable answer to the question should be as close as possible to the observations provided by each online user; 2) the higher the quality of the user, the more similar answers to the question will be provided [32], the loss function of model is:

$$\mathcal{L} = \sum_{i=1}^{M'} d\left(\theta; \bar{Z}, x_i\right) + \frac{1}{2}||w||^2, \tag{4}$$

where θ are all parameters in the model, Z is the final output of the model, $\bar{Z}$ is the mean of Z which is regarded as identified truth:

$$\bar{Z} = \frac{1}{M'}\sum_{i=1}^{M'} Z_i. \tag{5}$$

The weights $W^{(0)}$ and $W^{(1)}$ of neural network are trained using gradient decent. For our model, we perform batch gradient using the full notes in $\mathcal{G}'$ for every iteration. For sparse representation of A', the memory requirement is linear in the number of edges $\mathcal{O}(|\mathcal{E}'|)$, the computation complexity is liner in the number of graph edges $\mathcal{O}(|\mathcal{E}'|K^2H)$.

IV. Experiments

In this section, we present the results of the experiments that we conducted to validate the effectiveness of proposed model. We first introduce the dataset. We then present and analyze the experimental results. We also discuss the influence of different parameters on the results.

A. Implementation

In practice, we make use of TensorFlow for an efficient GPU-based implementation of proposed model using sparse-dense matrix multiplications.

B. Dataset

We use public dataset **Short Answer Scoring** to demonstrate the effectiveness of the proposed method. This dataset from Kaggle's competition The Hewlett Foundation: Short Answer Scoring. There are four subjects Science, English, English Language Arts and Biology in the dataset. Each of the dataset was generated from a question. All answers have an average length of 50 words per response which were written by students primarily in Grade 10 and scored by teachers from 0-3.

C. Experiment Protocols

1) Baseline Methods: We compare the proposed model with retrieval-based answer selection approaches and state-of-the-art truth discovery methods.

Retrieval-Based Answer Selection Approaches: the vectors of question and corresonding answers are extracted. Answers are ranked according to the similarities between the question vector and corresponding answer vectors. According to the difference of vector representation, it is specifically **Bag-of-Word (BOW) Similarity**, **Term Frequency-Inverse Document Frequency (TF-IDF) Similarity**, **Global Vectors for**

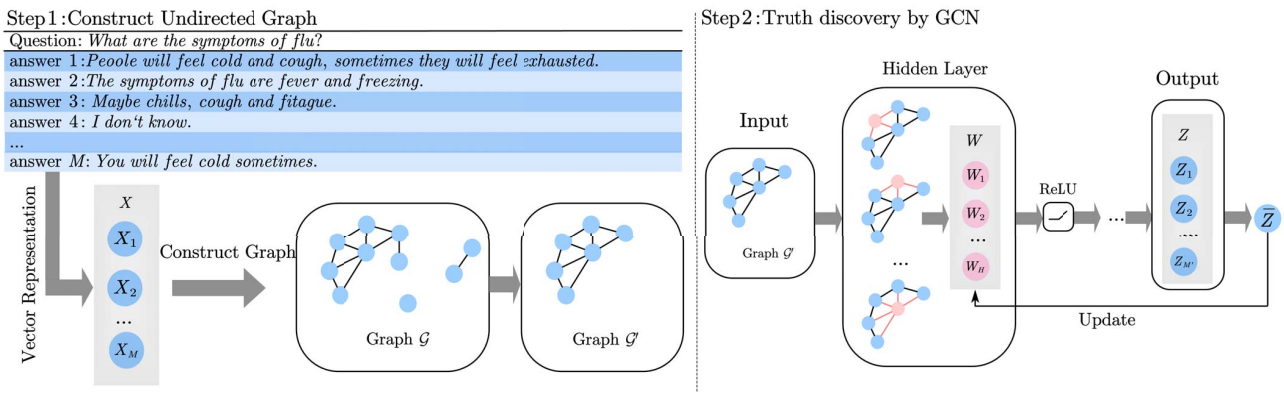

Fig. 1. Proposed Model

Word Representation (GloVe) Similarity and **SIF Similarity** four methods.

CRH [7]: To the best of our knowledge, CRH is regarded as relatively optimal truth discovery framework in recent years, which can handle both categorical and continuous data. We also use SIF to extract semantic information of answers. It should be noted that since the origin CRH method is not applicable to the text data, we also transforms the answers to vectors, and use the distance function proposed in this paper to measure the similarities between text answers. When using BoW and TF-IDF as vector representation approaches, since the text vectors are sparse, the result of truth discovery is a zero vector,, CRH method is invalid.

2) Evaluation Metrics: We returned the average score of Top-N(N is set as 10, 30, 50, 80, 100, 200, 300 in this paper) trustworthy answers from different methods of corresponding dataset.

D. Performance Comparison

The results are shown in Fig.2. As one can see, the proposed model consistently outperforms all the baseline approaches including retrieval-based approaches and state-of-art truth discovery approaches for all datasets. In other words, the proposed model demonstrates its great advantages on text data truth discovery. We analyzed the reasons for the superior performance of this model compared with retrieval-based approaches and state-of-the-art truth discovery approaches as follows:

Compared with retrieval-based approaches (i.e., BoW Similarity), which rank the answers merely based on the semantic similarity between the question and answers. However, a question itself may not cover all the semantics that should be covered in reliable answers. Therefore, retrieval-based methods only discover relevant answers rather than trustworthy answers.

Compared with state-of-the-art truth discovery approach CRH, although this method aims to capture user reliability, the performance is not great. It is because that this method assume the relationship between source reliability and answer credibility can be represented by simplified functions. This assumption leads to suboptimal text truth discovery results because the exact relational dependency between sources and claims is often unknown a priori. In addition, for scene with large number of sources but few observations, representation capability of real number as reliability of sources is limited. Different from baseline methods, our method use a GCN based model to learn the complex relationship and predict the reliable answers.

E. Parameter setting

We conducted a series of experiments to analyze the effect of parameters on the results of truth discovery from text data. We analyzed the following parameters, threshold α of similarity, threshold β of degree, and learning rate. These three parameters are considered to affect the performance of proposed model. Threshold α and β determine the graph constructed, and learning rate is an improtant parameter for proposed neural network.

1) Threshold α of Similarity: In the process of constructing graph, a threshold α is introduced to determine if two answer nodes are connected. Here we experimentally test the sensitivity of α. We try threshold α from 0.1 to 0.8. We find that α negligibly affects the results of the model. Due to space limitation, we only show the results on English dataset as Fig.3. The results on rest datasets follow the same tendency. According to our experiment experience, when the similarity α is defined as about 0.3, the results are relatively optimal. Meanwhile, we observe that parameter α can balance the information used with answer features or structure information of the graphs.

2) Threshold β of Degree: We fix the similarity of $\alpha = 0.3$, and try threshold β from 100 to 200. We find that β negligibly affects the results of the model too. Due to space limitation, we only show the results on English dataset as Fig.4. The results on rest datasets follow the same tendency. According to our experiment experience on all datasets, when the β is

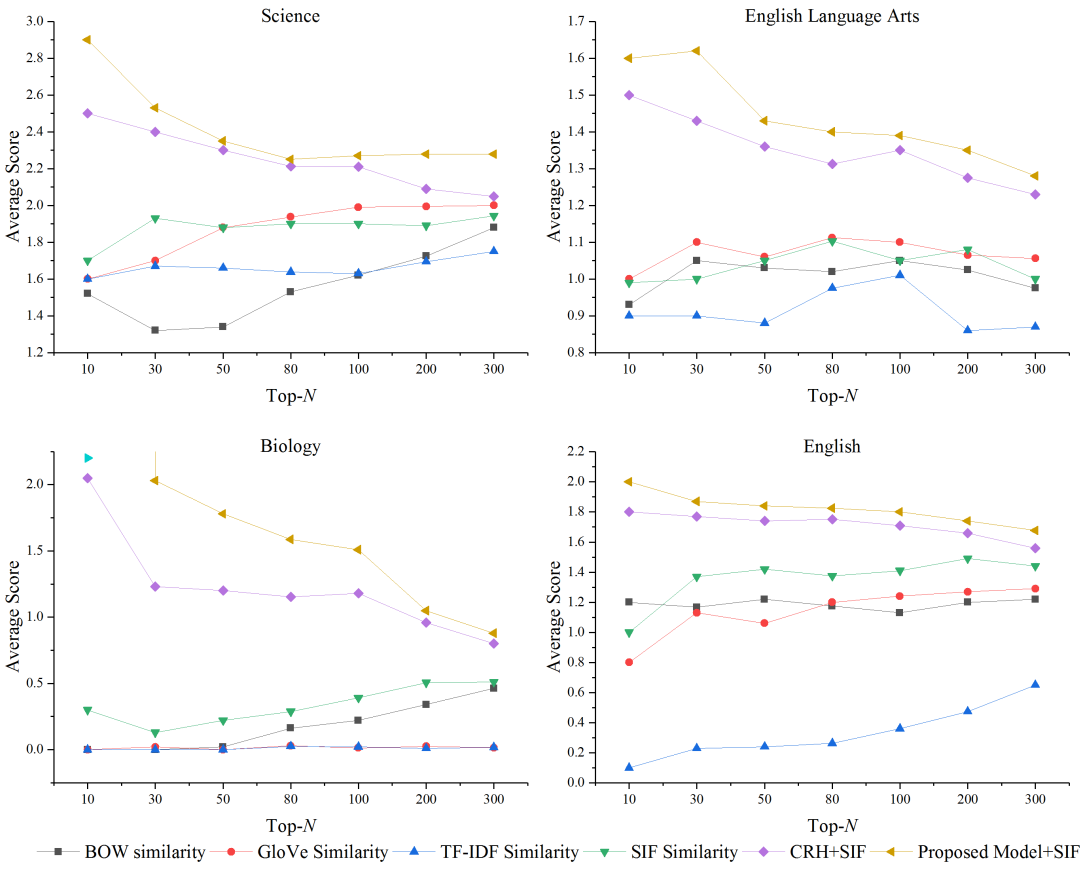

Fig. 2. Performance on Short Answer Scoring

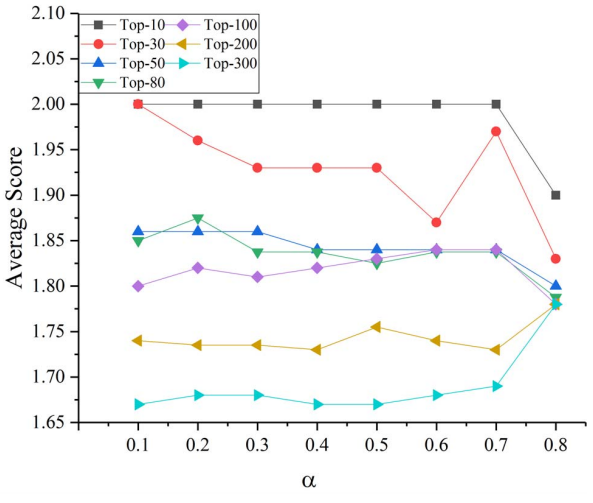

Fig. 3. Results Under Different α on Dataset English

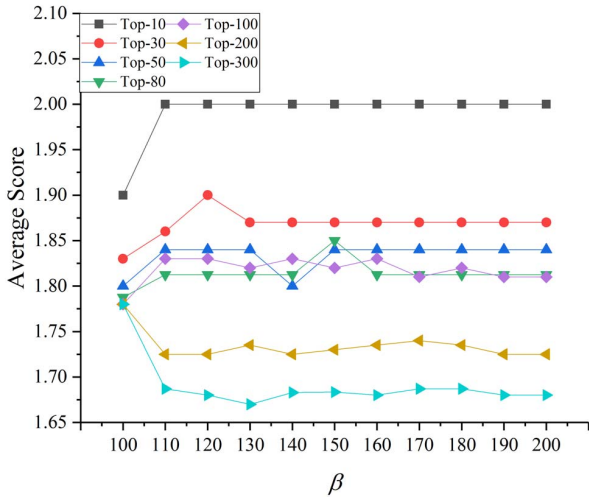

Fig. 4. Results Under Different β on Dataset English

defined as about 10 percent of the number of all answers, the results are relatively optimal.

3) Learning Rate: We try learning rates from 0.00001 to 0.1. As is shown in Fig. 5, we find that learning rate has limited effects on the results of the model. 0.001 is considered to be

a relatively optimal setting of learning rate.

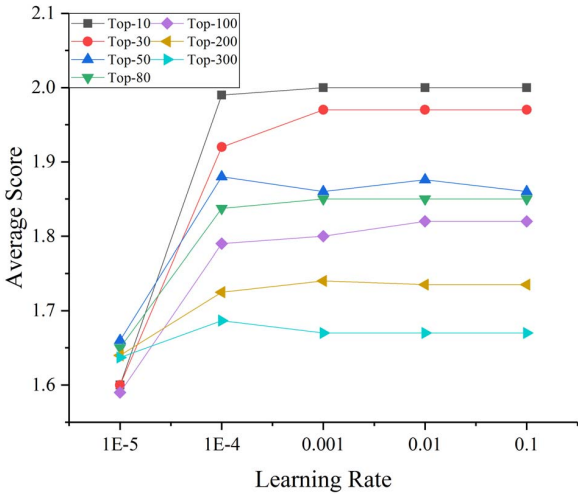

Fig. 5. Results Under Different Learning Rate on Dataset English

V. Conclusion

As an emerging topic, truth discovery has shown its effectiveness in a wide range of applications with structured data. However, existing methods all suffer on unstructured text data, and assume the relationship between source reliability and claim credibility can be represented by simplified functions. The structure information and semantic information are ignored during the truth discovery process. To tackle these challenges, in this paper, we propose a GCN based model that use graph of answers as input and outputs the ranking of answers based on the identified truth answer vector. More specifically, we first use graph convolutional network to sum up semantic information of all answers to assist the process of truth discovery. We utilize vectors to represent the reliability of answers, which have better representation capability than real numbers. Experimental results show that proposed model considerably outperforms the state-of-the-art methods. Proposed model can find trustworthy answers for each question in a unsupervised way.

Acknowledgment

This work was supported by National Science Foundation of China (Grant No.61371196), the China Postdoctoral Science Foundation Funded Project (Grand No.2009046425, 201003797).

References

[1] Y. Li, *et al.*, "A Survey on Truth Discovery." *ACM sigkdd Explorations Newsletter*, vol. 17, no.2, pp. 1-16

[2] A. Galland, *et al.*, "Corroborating information from disagreeing views," *ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 131-140, 2010.

[3] J. Pasternack, and R. Dan, "Knowing What to Believe(when you already know something)," *International Conference on Computational Linguistics (COLING)*, pp. 877-885, 2010.

[4] X. L. Dong, *et al.*, "Integrating Conflicting Data: The Role of Source Dependence," *Proceedings of the Vldb Endowment*, vol. 2, no.1, pp. 550-561, 2018.

[5] B. I. Aydin, *et al.*, "Crowdsourcing for multiple-choice question answering," *IAAI*, pp. 2946-2953, 2014.

[6] Q. Li, *et al.*, "A confidence-aware approach for truth discovery on long-tail data," *PVLDB*, vol. 8, no.4, pp. 425-436, 2014.

[7] Y. Li, *et al.*, "A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no.8, pp. 1986-1999, 2016.

[8] Y. Li, *et al.*, "On the Discovery of Evolving Truth," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 675-684, 2015.

[9] J. Pasternack, and R. Dan, "Latent credibility analysis," *the International Conference on World Wide Web (WWW'13)*, pp. 1009-1020, 2013.

[10] B. Zhao, *et al.*, "A Bayesian approach to discovering truth from conflicting sources for data integration," *PVLDB*, vol. 5, no.6, pp. 550-561, 2012.

[11] B. Zhao, and J. Han, "A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources," *Intl.workshop on Quality in Databases*,

[12] L. Li, *et al.*, "Truth discovery with memory network," *Tsinghua Science and Technology*, vol. 22, no.6, pp. 609-618, 2017.

[13] K. Broelemann, *et al.*, "Restricted Boltzmann Machines for Robust and Fast Latent Truth Discovery," *Proceedings of the Vldb Endowment*, 2017.

[14] R. Singh, *et al.*, "Neural Network Architecture for Credibility Assessment of Textual Claims," 2018.

[15] J. Marshall, *et al.*, "A Neural Network Approach for Truth Discovery in Social Sensing," *IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 343-347, 2017.

[16] K. Popat, *et al.*, "Credibility Assessment of Textual Claims on the Web," *25th ACM International on Conference on Information and Knowledge Management (CIKM)*, 2016.

[17] K. Broelemann, *et al.*, "LTD-RBM: Robust and Fast Latent Truth Discovery Using Restricted Boltzmann Machines," *?*, 2017.

[18] J. Marshall, *et al.*, "A Neural Network Approach for Truth Discovery in Social Sensing," *IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2017.

[19] Y. Li, *et al.*, "Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts," *ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

[20] H. Zhang, *et al.*, "TextTruth: An Unsupervised Approach to Discover Trustworthy Information from Multi-Sourced Text Data," *ACM SIGKDD International Conference on Knowledge Discovery (KDD)*, pp. 2729-2737, 2018.

[21] M. Bouguessa, *et al.*, "Identifying Authoritative Actors in Question-Answering Forums The Case of Yahoo! Answers," *SIGKDD*, 2008.

[22] L. Hong, and B. D. Davison, "A Classification-based Approach to Question Answering in Discussion Boards," *SIGIR*, 2009.

[23] Y. Liu, *et al.*, "CQARank: Jointly Model Topics and Expertise in Community Question Answering," *ACM International on Conference on Information and Knowledge Management (CIKM)*, 2013.

[24] G. Zhou, *et al.*, "Topic-sensitive probabilistic model for expert finding in question answer communities," *ACM International on Conference on Information and Knowledge Management (CIKM)*, 2012.

[25] Z. Wang, *et al.*, "Sentence Similarity Learning by Lexical Decomposition and Composition," 2017.

[26] X. Yao, *et al.*, "Answer Extraction as Sequence Tagging with Tree Edit Distance," *NAACL-HLT*,.PP.858867, 2013.

[27] M. Feng, *et al.*, "Applying Deep Learning to Answer Selection: A Study and An Open Task," *IEEE Workshop on ASRU.*, 2016.

[28] Y. Liu, *et al.*, "A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering," *ACL*, pp. 707-712, 2015.

[29] S. Arora, *et al.*, "A simple but tough-to-beat baseline for sentence embeddings," *Intl Conf. on Learning Representations (ICLR)*, 2017.

[30] S. Arora, *et al.*, "RAND-WALK: A Latent Variable Model Approach to Word Embeddings," *Computer Science*, pp. 1242-1250, 2015.

[31] T. N. Kipf, and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *Intl Conf. on Learning Representations (ICLR)*, 2016.

[32] R. Ma *et al.*, "MTruthsAn Approach of Multiple Truths Finding from Web Information," *Journal of Computer Research and Development*, 2016.

[33] M. Mohler *et al.*, "Text-to-text Semantic Similarity for Automatic Short Answer Grading," *Proceedings of the Conference of the European Association of Computational Linguistics*, 2009.