Table 9: The triples that learn similar truth vectors to (coronary heart disease, chest pain) with different embeddings.

Only Entity Embeddings		Only Co-occurrence Embeddings		Combine Two Kinds of Embeddings	
Similar Triple	Distance	Similar Triple	Distance	Similar Triple	Distance
(heart disease, chest pain)	0.9872	(heart disease, chest pain)	0.6546	(heart disease, chest pain)	0.9444
(myocardial infarction, chest pain)	1.0204	(cardio-cerebrovascular disease, chest pain)	0.7095	(cardio-cerebrovascular disease, chest pain)	1.0933
(coronary heart disease, left chest pain)	1.2282	(cardiovascular disease, coronary insufficiency)	0.7275	(coronary heart disease, dorsal distending pain)	1.2294
(coronary heart disease, chest distress)	1.2822	(myocardial infarction, filling defect)	0.7504	(heart disease, limited activity)	1.2520
(heart failure, chest pain)	1.4299	(coronary heart disease, shoulder pain)	0.8885	(myocardial infarction, filling defect)	1.2703

7 CONCLUSIONS

In this paper, we present a medical knowledge condition discovery method to enrich medical knowledge graph with condition information. Due to the limited amount of available EMR data, we leverage medical QA data from online crowdsourcing medical communities to overcome the lack of data. However, unlike EMR data, the quality of QA data is diverse, as the answers are provided by website users with different professional levels, which may introduce a lot of noise and degrade the quality of discovered conditions. To tackle these challenges, we propose a novel truth discovery method for the task of medical knowledge condition discovery. The proposed method can recognize the EMR data as priorly known high-quality reference sources to semi-supervise the overall process of medical knowledge condition discovery in multi-source medical data. Besides, the proposed method incorporates the occurrence and entity information of knowledge triples to capture the interaction between knowledge triples when computing the truth for knowledge triples. Experimental results on real-world medical datasets show that the proposed method can effectively discover accurate medical knowledge condition information from multi-source data with diverse quality. We also validate the effectiveness of the proposed method under various scenarios on synthetic datasets.

8 ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (No.61602013), the Shenzhen Fundamental Research Project (No.JCYJ20170818091546869), the Shenzhen Project (No.ZDSYS201802051831427), and the project "PCL Future Regional Network Facilities for Large-scale Experiments and Applications (PCL2018KP001)". Min Yang was sponsored by CCF-Tencent Open Research Fund.

REFERENCES

[1] Melisachew Wudage Chekol, Giuseppe Pirrò, Joerg Schoenfisch, and Heiner Stuckenschmidt. 2017. Marrying Uncertainty and Time in Knowledge Graphs.. In *AAAI*. 88–94.
[2] Yang Chen and Daisy Zhe Wang. 2014. Knowledge expansion over probabilistic knowledge bases. In *SIGMOD*. 649–660.
[3] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *SIGKDD*. 601–610.
[4] Xin Luna Dong, Barna Saha, and Divesh Srivastava. 2012. Less is more: selecting sources wisely for integration. In *VLDB*. 37–48.
[5] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing Ground Truth for Medical Relation Extraction. *TiiS* 8, 2 (2018), 11:1–11:20.
[6] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. 2011. YAGO2:exploring and querying world knowledge in time, space, context, and many languages. In *WWW*. 229–232.
[7] Abhyuday N Jagannatha and Hong Yu. 2016. Bidirectional RNN for Medical Event Detection in Electronic Health Records. In *NAACL-HLT*. 473.
[8] Jingchi Jiang, Chao Zhao, Yi Guan, and Qiubin Yu. 2017. Learning and inference in knowledge-based probabilistic model for medical diagnosis. *Knowledge-Based Systems* (2017).
[9] Charles Jochim and Lea Deleris. 2017. Named Entity Recognition in the Medical Domain with Constrained CRF Models. In *EACL*. 839–849.
[10] V Law, C Knox, Y Djoumbou, T Jewison, A. C. Guo, Y. Liu, A Maciejewski, D Arndt, M Wilson, and V Neveu. 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research* 42, Database issue (2014), 1091–7.
[11] J. Lehmann. 2015. DBpedia: A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web* 6, 2 (2015), 167–195.
[12] Cheng Li, Santu Rana, Dinh Phung, and Svetha Venkatesh. 2016. Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records. *Knowledge-Based Systems* 99, C (2016), 168–182.
[13] Qi Li, Yaliang Li, Jing Gao, Wei Fan, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD*. 1187–1198.
[14] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2015. Truth finding on the deep web: is the problem solved? *PVLDB* 6, 2 (2015), 97–108.
[15] Xian Li, Weiyi Meng, and C Yu. 2011. T-verifier: Verifying truthfulness of fact statements. In *ICDE*. 63–74.
[16] Yaliang Li, Nan Du, Chaochun Liu, Yusheng Xie, Wei Fan, Qi Li, Jing Gao, and Huan Sun. 2017. Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers from non-experts. In *WSDM*. 253–261.
[17] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Wei Fan, Wei Fan, Wei Fan, and Jiawei Han. 2016. A Survey on Truth Discovery. *Acm Sigkdd Explorations Newsletter* 17, 2 (2016), 1–16.
[18] Xueling Lin and Lei Chen. 2018. Domain-Aware Multi-Truth Discovery from Conflicting Sources. *PVLDB* 11, 5 (2018), 635–647.
[19] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, and Divesh Srivastava. 2011. Online Data Fusion. *PVLDB* 4, 11 (2011), 932–943.
[20] Shanshan Lyu, Wentao Ouyang, Huawei Shen, and Xueqi Cheng. 2017. Truth Discovery by Claim and Source Embedding. In *CIKM*. 2183–2186.
[21] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *SIGKDD*. 745–754.
[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Computer Science* (2013).
[23] Changsung Moon, Paul Jones, and Nagiza F. Samatova. 2017. Learning Entity Type Embeddings for Knowledge Graph Completion. In *CIKM*. 2215–2218.
[24] Jingchao Ni, Hongliang Fei, Wei Fan, and Xiang Zhang. 2017. Automated Medical Diagnosis by Ranking Clusters Across the Symptom-Disease Network. In *ICDM*. 1009–1014.
[25] Jeff Pasternack and Dan Roth. 2010. Knowing what to believe (when you already know something). In *COLING*. 877–885.
[26] Ying Shen, Yang Deng, Jin Zhang, Yaliang Li, Nan Du, Wei Fan, Min Yang, and Kai Lei. 2018. IDDAT: An Ontology-Driven Decision Support System for Infectious Disease Diagnosis and Therapy. In *ICDM Workshops*. 1417–1422.
[27] Julien Tourille, Olivier Ferret, Aurelie Neveol, and Xavier Tannier. 2017. Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers. In *ACL*. 224–230.
[28] Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. 2018. Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning with Confidence. In *AAAI*.
[29] Hao Xin, Rui Meng, and Lei Chen. 2018. Subjective Knowledge Base Construction Powered By Crowdsourcing and Knowledge Base. In *SIGMOD*. 1349–1361.
[30] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2007. Truth discovery with multiple conflicting information providers on the web. In *SIGKDD*. 1048–1052.
[31] Xiaoxin Yin and Wenzhao Tan. 2011. Semi-supervised truth discovery. In *WWW*. 217–226.
[32] Q. Yuan, J. Gao, D. Wu, S. Zhang, H Mamitsuka, and S. Zhu. 2016. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32, 12 (2016), i18–i27.
[33] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2018. On the Generative Discovery of Structured Medical Knowledge. In *SIGKDD*. 2720–2728.