Tue. Jun 4, 2019

Room K

International Session | International Session | [ES] E-1 Knowledge engineering

[1K4-E-1] Knowledge engineering

Chair: Kazushi Okamoto (The University of Electro-Communications), Reviewer: Yasufumi Takama (Tokyo Metropolitan University)
5:20 PM - 7:00 PM Room K (201A Medium meeting room)

[1K4-E-1-01] How is Social Capital Associated with Perception of AI?
○Yoji Inaba[1] （1. Nihon University）
5:20 PM - 5:40 PM

[1K4-E-1-02] Multimodal Neural Network– based Health Platform for Knowledge Decision-Making
○Kyungyong Chung[1] （1. Kyonggi University）
5:40 PM - 6:00 PM

[1K4-E-1-03] Variables Extraction in Natural (English) Language Through Possessive Relationships
○Danilo Eidy Miura[1], Teruaki Hayashi[1], Yukio Ohsawa[1] （1. The University of Tokyo）
6:00 PM - 6:20 PM

[1K4-E-1-04] Using Sequence Constraints for Modelling Network Interactions
Johannes De Smedt[2], ○Junichiro Mori[1], Masanao Ochi[1] （1. The University of Tokyo, 2. The University of Edinburgh）
6:20 PM - 6:40 PM

[1K4-E-1-05] CTransE : Confidence-Based Translation Model for Uncertain Knowledge Graph Embedding
○Natthawut Kertkeidkachorn[1,2], Xin Liu[1], Ryutaro Ichise[2,1] （1. National Institute of Advanced Industrial Science and Technology, 2. National Institute of Informatics）
6:40 PM - 7:00 PM

©The Japanese Society for Artificial Intelligence

International Session | International Session | [ES] E-1 Knowledge engineering

[1K4-E-1] Knowledge engineering

Chair: Kazushi Okamoto (The University of Electro-Communications), Reviewer: Yasufumi Takama (Tokyo Metropolitan University)

Tue. Jun 4, 2019 5:20 PM - 7:00 PM Room K (201A Medium meeting room)

[1K4-E-1-01] How is Social Capital Associated with Perception of AI?

○Yoji Inaba[1] （1. Nihon University）

5:20 PM - 5:40 PM

[1K4-E-1-02] Multimodal Neural Network– based Health Platform for Knowledge Decision-Making

○Kyungyong Chung[1] （1. Kyonggi University）

5:40 PM - 6:00 PM

[1K4-E-1-03] Variables Extraction in Natural (English) Language Through Possessive Relationships

○Danilo Eidy Miura[1], Teruaki Hayashi[1], Yukio Ohsawa[1] （1. The University of Tokyo）

6:00 PM - 6:20 PM

[1K4-E-1-04] Using Sequence Constraints for Modelling Network Interactions

Johannes De Smedt[2], ○Junichiro Mori[1], Masanao Ochi[1] （1. The University of Tokyo, 2. The University of Edinburgh）

6:20 PM - 6:40 PM

[1K4-E-1-05] CTransE : Confidence-Based Translation Model for Uncertain Knowledge Graph Embedding

○Natthawut Kertkeidkachorn[1,2], Xin Liu[1], Ryutaro Ichise[2,1] （1. National Institute of Advanced Industrial Science and Technology, 2. National Institute of Informatics）

6:40 PM - 7:00 PM

©The Japanese Society for Artificial Intelligence

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

How is Social Capital Associated with Perception of AI? – An Observation from a Survey of Residents in Metropolitan Tokyo Area

Yoji Inaba[*]

[*1] Nihon university

Abstract

Numerous studies over the past 30 years have examined the relationship between social capital (SC) and information and communication technology (ICT). However, few studies have examined the association between artificial intelligence (AI) and SC. This study addresses this gap using a Web survey (N=5000) carried out in the Tokyo metropolitan area in Japan in 2018. The survey included questions on ICT literacy and SC (networks, trust, norms of reciprocity), as well as questions on perceptions of AI including its impact on society. The author found a statistically significant positive association between cognitive SC (trust and norms of reciprocity) and positive perceptions of AI. However, the impact of structural SC (networks) on AI perceptions was either nonexistent or negative. Structural SC created by group participation, as well as contact with others, including those in the workplace, does not create positive perceptions of AI, and could even be a source of negative perceptions of AI. Cognitive SC may function as a promoter of AI, while structural SC may function as a precaution to AI. Both types of SC might assume important roles for the smooth transition to the AI era.

Keyword: Social Capital, ICT, AI, Networks

1. Introduction

1. Introduction

Numerous papers and articles have been written on the relation between social capital (SC) and information communication technologies (ICT) in the past thirty years. Artificial intelligence (AI) which obviously overlaps with ICT is another popular subject in recent years. Yet few papers deal with the association between AI and SC. This paper is an attempt to fulfill the gap based on a web survey (N=5000) the author carried out in the metropolitan Tokyo area in Japan in 2018. The survey asked questions on ICT literacy, SC (networks, trust, norms of reciprocity) as well as perceptions on the AI and its impact on society.

2. Preceding studies and research questions

2.1 Preceding studies

Introduction of ICT until the middle of 2000s functioned both complementary and/or substitutionary to the existing offline SC (1st Wave). Then online structural SC created by ICT mainly enhanced offline cognitive SC (2nd Wave). Currently the offline SC backed up by online structural SC mostly seems to have positive impact on ICT applications to the society. Although ICT literacy and SC were originally mutually correlated, main concern on causality have shifted from ICT to SC rather than vice versa.

2.2 Research questions

The present study deals with the following research questions which have not been fully fulfilled by preceding studies.

RQ1 How will SC affect ICT literacy? Although many papers have dealt with ICT, most of them examined impact of just a particular type of ICT. The lack of comprehensiveness could be said about the preceding studies dealing with SC. As for SC, it covers various concepts including networks, trust, and norms. To the best of our knowledge, there is no paper which is based upon comprehensive pictures of both ICT literacy and SC. This is especially the case on the impact of SC to the ICT literacy. Therefore the present paper shed the light on this aspect.

RQ2 How will SC affect the perception of AI? Are there any implications which could be inferred by the way SC is associated with ICT applications?

RQ3 How will AI affect SC? Are there any implications which could be inferred from an analysis on the way ICT literacy is associated with SC?

Contact: Yoji Inaba, Nihon Univesity, Phone +81-03-5275-8636, Fax +81-03-5936-1780, inaba.yoji@nihon-u.ac.jp

RQ4 What are the policy implications to cope with negative impact of AI introduction if there is any?

3. Materials and Methods

3.1 Data

The dataset used in this study comes from a questionnaire survey that we conducted between September 4th and 10th, 2018 over the internet. The survey covers residents in Metropolitan Tokyo area including Tokyo, Kanagawa, Chiba, and Saitama prefectures between the ages of 20 and 69 years. The survey questionnaires were sent and recovered through the internet. We got reply from 5000 people.

The Questionnaire survey was conducted on the following topics:

① ICT literacy

・8 questions on the availability of ICT devices (yes or no),

・14 questions on the frequency of use of ICT devices and web services with a three-point Likert scale,

・8 questions on capability of soft wares and web services usage with a four-point Likert scale,

・6 questions on the experience of using AI related equipment with a three-point Likert scale.

② SC

12 questions are asked with regard to SC.

・8 questions related to structural SC including group participation, relationships with neighbors, family members and relatives, friends and acquaintances, and colleagues (from four to seven-point Likert scales),

・4 questions related to cognitive SC including trust and norm of reciprocity with a four-point Likert scale.

② Perception on AI

The questionnaire dealt with 30 items on respondents' perception with regard to AI.

・7questions about evaluations on the influence of AI (one five-choice question and 6 Likert scale questions with a four-point)

・8 questions related to pros & cons on AI use in our society at 8 situations with a five-point Likert scale,

・8 questions on personal preference of AI use at 8 situations with a five-point Likert scale,

・7 questions on the choice between AI or human beings at 7 situations with a four-point Likert scale.

In addition, the survey asked personal attributes of respondents including sex, age, educational attainment, marital status, occupation, forms of employment (permanent or temporary), family income, number of cohabiting people, duration of residence, the name of municipality they reside, and perception on risk.

Concerning the ethical appropriateness of the contents of the questionnaire, the survey was checked and approved by the ethical committee (social science) of the Tohoku University.

3.2 Methodology

I conducted three factor analyses using the above mentioned data on ICT literacy, SC, and perception of AI in order to get basic factors out of these questions in each of the three groups. Then, using factor scores as explanatory variables, two logistic regressions are carried out. First, logistic regressions between ICT literacy and SC. As an extension of the third wave of preceding studies which try to explain the behavior of ICT by SC, we used SC factors as independent variables to explain the changes in ICT literacy factors. Secondly, another series of logistic regression on perceptions of AI, using SC as independent variable controlling ICT literacy and characteristics of respondents.

4. Results

Statistically significant positive associations between cognitive SC and affirmative perception of AI use were observed. However, impact of structural SC on AI perception is either none existent or negative. Structural SC created by group participation as well as contacts with others including those at work place does not create affirmative perception on AI. On the contrary, they could be a source of negative AI perception.

References

Ahmed, Zafor (2018) "Explaining the unpredictability : a social capital perspective on ICT intervention", International Journal of Information Management,38, 175-186.

Bauernschuster.S, F.Oliver, W. Ludger,(2014) "Surfing alone? The internet and social capital: Evidence from an unforeseeable technological mistake", Journal of Public Economics, vol. 117, issue C,73-89

Terashima. K, and A. Miura(2013) "Does use of SNS develop offline/online social caital", Kwansei Gakuin University Bulletin of Psycological Science, Vol.39,59-67. (in Japanese)

Kobayashi, T., & Ikeda, K.（2006）. The development of social capital in communities in an online game: A perspective o-n a"spill over"effect into the offline world. Japanese Journal of Social Psychology, vol.22-1, 58−71.(in Japanese）

Hsieh, J.J.P., Rai, A., & Keil, M. (2010). Addressing digital inequality for the socio- economically disadvantaged through government initiatives; Forms of capital that affect ICT utilization. Information Systems Research, 233-253.

Ishizuka, M, et al. (2017) "Introduction" In The Japan Society for Artificial Intelligence (ed.) Encyclopedia of Artificial Intelligence, Kyouritsu Publishing, p.2 (in Japanese) .

Li, Xiaoqian and Wenhong Chen (2014) "Facebook or Renren? A comparative study of social networking site use and social capital among Chinese international students in the United States" Computers in Human Behavior 35,116–123

Lu, J., Yang, J., & Yu, C. (2013). Is social capital effective for online learning? Information & Management, 50, 507-522.

Miyakawa, Tadao, and Omori Takashi (2004) Social Capital, Toyo Keizai (in Japanese).

Miyta, Kakuko (2005a) Social Phycology of the Internet – the function of the internet from a viewpoint of social capital, Kazama Syobo (in Japanese).

Miyata, Kakuko(2005b) Media as a bridge among human ties – Social Capital in the Era of the Internet, NTT Publishing (in Japanese).

Miyata, Kakuko, Barry Wellman, and Jeffrey Boase. (2005c) "The Wired — and Wireless — Japanese: Webphones, PCs and Social Networks." Mobile Communications. Computer Supported Cooperative Work, Springer, vol 31. pp.427-449.

Naranjo-Zolotov, Mijail et al. (2019) "Examining social capital and individual motivators to explain the adoption of online citizen participation", Future Generation Computer Systems,
92 302–311

Nie, Norman, and Lutz Erbring (2002) "Internet and society: Preliminary Report", IT and SOCIETY, Vol 1, 275-283.

Nolan.S, Hendricks.J, Amanda.T,(2015) " Social networking sites (SNS); exploring their uses and associated value for adolescent mothers in Western Australia in terms of social support provision and building social capital" Midwifery 31,912–919

Norris,Pippa (2003),"Social Capital and ICTs: Widening or Reinforcing Social Networks?"presented at the"International Forum on Social Capital for Economic Revival"held by the Economic and Social Research Institute, Cabinet Office, Japan in Tokyo,24-25th March 2003、

Penard,Thierry and Nicolas Poussing (2010) Internet use and social capital; the strength of virtual ties, Journal of Economic Issues, 44(3), 569-595.

Ryan, S. (2010). Information system and healthcare XXXVI: Building and maintaining social Capital-Evidence from the field. Communications of the Association of Information Systems, 27(18), 307-322.

Salahuddin, Mohammad et al. (2016) "Does internet stimulate the accumulation of social capital? A macro-persective from Australia" Economic Analysis and Policy, 49, 43-55.

Zhong, Zhi-Jin (2014) "Civic engagement among educated Chinese youth: The role of SNS (Social Networking Services), bonding and bridging social capital" Computers & Education 75,263–273urname of the first author Year] Names of authors, Title, Journal, Publisher, Year.

Multimodal Neural Network-based Health Platform for Knowledge Decision-Making

Joo-Chung Kim[1], Ji-Won Baek[2], Hyun Yoo[3], Kyungyong Chung[*4]

[1,2]Data Mining Lab., Dept. of Computer Science, Kyonggi University, South Korea

[3]Dept. of Computer Information Engineering, Sangji University, South Korea

[*4]Division of Computer Science and Engineering, Kyonggi University, South Korea

There is a need for artificial intelligence-oriented information technologies (aimed at continuous monitoring and life-care of chronic diseases through health platforms) that can discover potential health-risk-factor changes and predict emerging risks. In this paper, we propose a multimodal neural network-based health platform for knowledge decision-making. The proposed method learns the relationships present between heterogeneous data and the multimodal neural network, and extracts the common information shared between the modals to estimate health-risk factors. The correlation of variables appearing in the health platform is used to construct a multimodal neural network, and shared common information is combined to estimate the health-risk factors. The correlations of the variables are shown as positive correlations and negative correlations. A positive correlation indicates a relationship in which two variables change in the same direction, and a negative correlation indicates a relationship in which they change in a different direction. The proposed multimodal neural network is used to solve the health-risk–factor problem in the health platform, improving the reliability of the data.

1. Introduction

Due to the development of information-gathering technology, a variety of data is gathered from various fields, such as society, science, and industry, and is being accumulated as big data (Rho et al., 2015). This is characterized by a continuous increase in volume, rapid change, and various properties (Chung and Roy, 2016). Especially in the health platform, the scope is expanding with the development of information processing and collection technologies such as the electronic medical record (EMR), personal health record (PHR), and life-log (Kim and Chung, 2017). In a health platform, data are mainly continuous, changing over time. Data related to human health are affected by internal and external environments. Health status changes over a short period of time or over a long period of time from variables such as weather, physical information, nutrition, and activity, etc. (Larose et al., 2014; Kim and Chung, 2018; Yoo and Chung, 2018). In these health data, the collection range changes depending on the user's surrounding environment or the device held. Missing values are likely to be generated in collected data due to differences in environments or devices among users (Sarwar et al., 2001; Chung et al., 2016). In particular, devices using low power have difficulty collecting real-time data, leading to missing values. In addition, even if the same type of device is used, the obtained data may show different properties.

Previous research (Kim and Chung, 2018) presented data mining of health-risk factors using the PHR, the EMR, and health status similarities from medical big data in a hybrid peer-to-peer (P2P) network environment. This data mining is used to provide a health service and a user-oriented healthcare promotion service for chronic diseases requiring constant care, life care, and elderly health care in health platforms. In addition, the present study predicts the potential health status of chronic diseases using a similarity-based sequence-mining algorithm.

This study is organized as follows: Section 2 describes the related researches of the health-risk factors of the health matrix, Section 3 describes the proposed multimodal neural network-based health platform for knowledge decision-making, Section 4 describes the multimodal based ontology mining for an adaptive knowledge processing, and Section 5 provides a conclusion

2. Related Works

Medical big data have a health-risk–factor problem in which the values of certain properties are missing due to human, natural, and mechanical errors in the collection process (Kim et al., 2017). Health-risk factors create a null state where there is no value for a variable or property at a particular point in time (Kim et al., 2014; Chung and Lee, 2004). Health factors are an important issue in data analysis and utilization, and various studies are under way to solve this problem. In addition, a higher percentage of missing values causes problems in the reliability of the results from a data analysis (Jung and Chung, 2016; Yoo and Chung, 2018; Phanich et al., 2010).

Health-risk–factor processing mainly uses value estimation or a substitution method, and if the influence of a property causing a large amount of health-risk factors is small, the property itself is removed. There are various methods to deal with health-risk factors: mean/median value substitution, the artificial neural network, the regression model, k-nearest neighbors, and collaborative filtering. Of the various methods, artificial neural networks have received much attention in recent years. This is an issue in various fields, such as object recognition, classification, and artificial intelligence, as well as in health-risk–factor estimation (EI-Dosuky et al., 2010; Jung and Chung, 2016; Kim

Contact: Kyungyong Chung, Kyonggi University, South Korea,
82-10-6362-4555, dragonhci@gmail.com

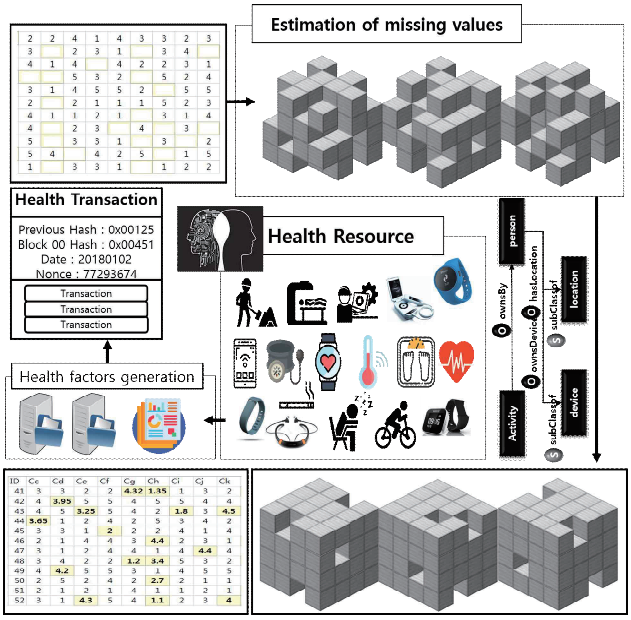
et al., 2014). In an artificial neural network, nodes and weights can be flexibly configured according to data characteristics.

3. Multimodal Neural Network–based Health Platform for Knowledge Decision-Making

3.1 Data Features in Health Platforms

Health data can be collected based on the same time, and they show a time-series characteristic that varies with time. An analysis of this makes it possible to distinguish whether the variables are positively or negatively correlated (Specht, 1993; Deshpande and Karypis, 2004; Chung et al., 2018). The closer the variables are to -1, the higher the negative correlation, and the closer to +1, the higher the positive correlation (Orciuoli and Parente, 2017; Adomavicius and Tuzhilin, 2015). Figure 1 shows the process of health-risk factors in health platforms.

In addition, being closer to 0 indicates that the two variables have no effect on each other. In this paper, health-risk factors are estimated through the neural network configuration using the correlations between variables.



2	2	4	1	4	3	3	2	3
3		2	3	1		3	4	
4	1	4		4	2	2	3	1
		5	3	2		5	2	4
3	1	4	5	5	2		5	5
2		2	1	1	1	5	2	3
4	1	1	2	1		3	1	4
4		2	3		4		3	
5		3	3	1		3		2
5	4		4	2	5		1	5
1		3	3	1		1	2	2

ID	Cc	Cd	Ce	Cf	Cg	Ch	Ci	Cj	Ck
41	3	3	2	2	**4.32**	**1.35**	1	3	2
42	4	**3.95**	5	5	4	5	5	4	4
43	4	5	**3.25**	5	4	2	**1.8**	3	**4.5**
44	**3.65**	1	2	4	2	5	3	4	2
45	3	3	1	**2**	2	2	4	1	4
46	2	1	4	4	3	**4.4**	2	3	1
47	3	1	2	4	4	1	4	**4.4**	4
48	3	4	2	2	**1.2**	**3.4**	5	3	2
49	4	**4.2**	5	5	3	1	4	5	5
50	2	5	2	4	2	**2.7**	2	1	1
51	2	1	2	1	4	1	1	2	1
52	3	1	**4.3**	5	4	**1.1**	2	3	**4**

Figure 1 Process of health-risk factors in health platform

3.2 Mining Health-risk Factors using a Multimodal Neural Network

This is a correlation-based neural network for estimating health-risk factors in the platform. Figure 2 shows the estimation of health-risk factors using a multimodal neural network in a health platform.

Positive correlations and negative correlations between the variables are analyzed, and the variables representing positive correlations are grouped into neural networks. According to the results of the analysis, a fully connected network is constructed with input values by grouping the variables with a high positive correlation. This makes it possible to estimate the health-risk factors appearing in the health data. The relationship between multimodals is learned from large-capacity heterogeneous data, and common information shared between modals is extracted from multimodal input based on it.

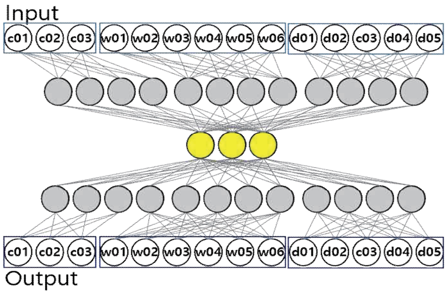


Figure 2 Process of health-risk factors in health platform

4. Multimodal based Ontology Mining for Adaptive Knowledge Processing

4.1 Adaptive Ontological Knowledge Representation

Health platforms use ontology-reasoning engines to express knowledge based on natural language. Mining is used to discover rules and to create knowledge using the discovered reasoning rules. In addition, knowledge is expanded by using inference engines according to the changing situations. Ontological knowledge representation specifies the relationship between data and attributes, and creates adaptive knowledge using language-based sequence tagging models. It consists of a top-level ontology for real-time service, and expansion in the previously developed health ontology (Chen and Tsai, 2018; Orciuoli and Parente, 2017). Figure 3 shows an adaptive ontological knowledge representation using ontology-reasoning engines.

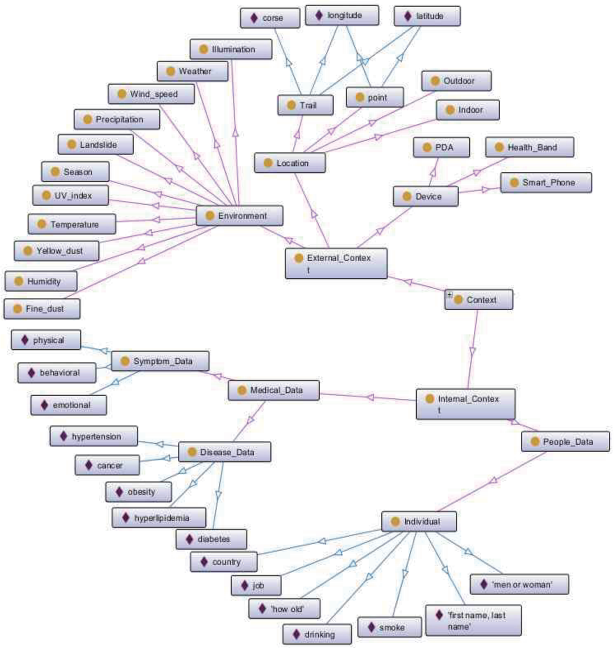


Figure 3 Adaptive ontological knowledge representation

The mining model is derived by using optimal tag prediction methods, using tag accuracy that is generated repeatedly by acquiring and refining natural language–based knowledge. An ontological knowledge model is also used to develop knowledge representation, knowledge-type conversion, and generation technologies to deal with learning data in three dimensions (Adomavicius and Tuzhilin, 2015). This can resolve the problem of a shortage in learning data by increasing the utility of learning and the use of knowledge. Knowledge is efficiently integrated and managed by configuring mining models according to predefined learning models. The model is configured to improve time, cost, and the integrity of knowledge management.

4.2 Multimodal-based Ontology Mining in Health Platforms

In the health platform, a mining model discovers association rules by constructing transactions from ontological knowledge. The life-log is collected in real-time through ambient sensors and is processed using multimodal knowledge acquisition technologies. In order to acquire knowledge based on natural language, a model is developed to extract and refine data that are considered knowledge from unstructured, semi-structured, and structured data. In the case of incomplete knowledge about the acquired information, characteristics that have not been acquired from the knowledge are inferred through mining using deep learning (Kim et al., 2014; Kim and Chung, 2017). Behavioral tips or potential risks are provided based on the inference rules related to healthcare. Users can avoid or prevent health risks by reflecting information they receive from making decisions based on behavioral predictions in their daily lives. In addition, changes in time series data are predicted through mining to provide warnings and countermeasures to users. Figure 4 shows the multimodal-based ontology mining process of health data sources, knowledge collection, knowledge processing, and knowledge analysis in health platforms.

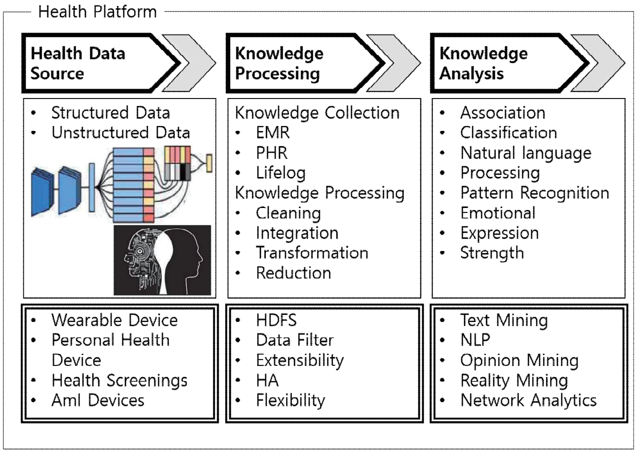


Figure 4 Multimodal-based ontology mining process

5. Conclusion

In this paper, we proposed a multimodal neural network-based health platform for knowledge decision-making. This is solved by a multimodal neural network that extracts common information about health-risk factors from multiple heterogeneous devices. In order to reflect the dynamic characteristics of the time series data, the variables showing a high correlation are grouped into multimodal neural networks for each cluster, and the health-risk factors are estimated by integrating them. The correlations between the variables were positively correlated with negative correlations. By using the proposed multimodal neural network, it is possible to construct the data with higher reliability in the health platform when estimating health-risk factors.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2016R1D1A1A09917313)

References

[1] M. J. Rho, K. S. Jang, K. Y. Chung, I. Y. Choi, "Comparison of Knowledge, Attitudes, and Trust for the Use of Personal Health Information in Clinical Research", Multimedia Tools and Applications, Vol. 74, No. 7, pp. 2391-2404, 2015.

[2] K. Chung, Roy C. Park, "PHR Open Platform based Smart Health Service using Distributed Object Group Framework", Cluster Computing, Vol. 19, No. 1, pp. 505-517, 2016.

[3] J. C. Kim, K. Chung, "Depression Index Service using Knowledge based Crowdsourcing in Smart Health", Wireless Personal Communication, Vol. 93, No. 1, pp. 255-268, 2017.

[4] D. T. Larose, C. D. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining", John Wiley & Sons, 2014.

[5] J. C. Kim, K. Chung, "Mining Health-Risk Factors using PHR Similarity in a Hybrid P2P Network", Peer-to-Peer Networking and Applications, Vol. 11, No. 6, pp. 1278-1287, 2018.

[6] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Item-based collaborative filtering recommendation algorithms", In Proc. of the 10th international conference on World Wide Web, pp. 285-295, 2001.

[7] K. Chung, J. C. Kim, R. C. Park, "Knowledge-based Health Service considering User Convenience using Hybrid Wi-Fi P2P", Information Technology and Management, Vol. 17, No. 1, pp. 67-80, 2016.

[8] J. C. Kim, K. Chung, "Depression Index Service using Knowledge based Crowdsourcing in Smart Health", Wireless Personal Communication, Vol. 93, No. 1, pp. 255-268, 2017.

[9] J. H. Kim, D. Lee, K. Y. Chung, "Item Recommendation based on Context-aware Model for Personalized u-Healthcare Service", Multimedia Tools and Applications, Vol. 71, No. 2, pp. 855-872, 2014.

[10] K. Y. Chung, J. H. Lee, "User Preference Mining through Hybrid Collaborative Filtering and Content-based Filtering in Recommendation System", IEICE Transaction on Information and Systems, Vol. E87-D, No. 12, pp. 2781-2790, 2004.

[11] H. Jung, K. Chung, "Knowledge-based dietary nutrition recommendation for obese management", Information Technology and Management, Vol. 17, No. 1, pp. 29-42, 2016.

[12] H. Yoo, K. Chung, "Mining-based Lifecare Recommendation using Peer-to-Peer Dataset and Adaptive Decision Feedback", Peer-to-Peer Networking and Applications, Vol. 11, No. 6, pp. 1309-1320, 2018.

[13] M. Phanich, P. Pholkul, S. Phimoltares, "Food Recommendation System Using Clustering Analysis for

Diabetic Patients", In Proc. of the International Conference on Information Science and Applications, pp. 1-8. 2010.

[14] M. A. El-Dosuky, M. Z. Rashad, T. T. Hamza, A. H. El-Bassiouny, "Food Recommendation Using Ontology and Heuristics", In Proc. of the International Conference on Advanced Machine Learning Technologies and Applications, pp. 423-429, 2012.

[15] H. Jung, K. Chung, "Life Style Improvement Mobile Service for High Risk Chronic Disease based on PHR Platform", Cluster Computing, Vol. 19, No. 2, pp. 967-977, 2016.

[16] J. H. Kim, J. Kim, D. Lee, K. Y. Chung, "Ontology Driven Interactive Healthcare with Wearable Sensors", Multimedia Tools and Applications, Vol. 71, No. 2, pp. 827-841, 2014.

[17] D. F. Specht, "The General Regression Neural Network-rediscovered. Neural Networks", Vol. 6, No. 7, pp. 1033-1034. 1993.

[18] M. Deshpande, G. Karypis, "Item-based top-n recommendation algorithms", ACM Transactions on Information Systems (TOIS), Vol. 22, No. 1, pp. 143-177, 2004.

[19] K. Chung, H. Yoo, D. E. Choe, "Ambient Context-based Modeling for Health Risk Assessment Using Deep Neural Network", Journal of Ambient Intelligence and Humanized Computing, 2018. DOI: 10.1007/s12652-018-1033-7

[20] T. Chen, H. R. Tsai, "Application of industrial engineering concepts and techniques to ambient intelligence: a case study", Journal of Ambient Intelligence and Humanized Computing, Vol 9, No. 2, pp. 215-223, 2018.

[21] F. Orciuoli, M. Parente, "An ontology-driven context-aware recommender system for indoor shopping based on cellular automata", Journal of Ambient Intelligence and Humanized Computing, Vol 8, No. 6, pp. 937-955, 2017.

[22] G. Adomavicius, A. Tuzhilin, "Context-Aware Recommender Systems", Recommender Systems Handbook pp 191-226, 2015.

1K4-E-1-03

Variables Extraction in Natural (English) Language Through Possessive Relationships

Danilo Eidy Miura
The University of Tokyo

Teruaki Hayashi
The University of Tokyo

Yukio Ohsawa
The University of Tokyo

The already highlighted importance of the 'flow' of data in the Market of Data brings needs of development of ways to better explore the utilization of data. Aware of the existence of rich knowledge stored and shared in text format, this paper aims to propose a form of representation of variable names that can be identified in natural language written knowledge. With the use of possessive relationships between words in Noun Phrases, we supported the representation of variable name relating a variable to a thing or event. A simple experiment was performed to demonstrate the efficacy of the proposed representation supported by Data Jacket Store, where we can find well-form variable names under the name of Variable Labels.

1. Introduction

The Chance Discovery in the Market of Data is a field of research that aims to design the flow of data in the society through creative methods and find hidden patterns. From the generation, through supply, processing, distribution to utilization of data, discoveries can enhance the demands and pull the consumption of data in the society. Therefore, the perception of the value of data through its utilization is the essential force to innovate in market. Although data is offered and advertised in online repositories, such as Data Jacket Store (Hayashi & Ohsawa, 2015), potential users of data may not be aware of the potential utilization of the offered data. Knowledge of data analysis and processing is required to explore the applications of data.

In order to support data users or suppliers to assess the value of their data, the exploration of potential utilization is the basis for valuing what is not in use yet. To support users in the exploration of potential use of data, we aim to recognize potential use of data from repositories of knowledge recorded in texts, such as research papers and data analytics reports. In this paper we explain the formal representation of data as a well formed variable name (WFVN) that can be used to discover potential new variable labels to name data.

In the field of Knowledge Engineering, the recognition of variables is an essential step to design the knowledge-based system. Implemented systems will contain well-formed variables and knowledge representation. But in natural language, the variables may be expressed in free style, enabling a direct codification to computer-readable language. In order to get the advantage of accumulated knowledge written in natural language, the variable identification and its formal representation may allow knowledge-based systems users to explore possibilities of data utilization.

2. Related Work

2.1.Knowledge Representation

In previous works in knowledge representation (Studer et al., 1998, Davis et al., 1993), we learned that the representation of the knowledge depends on the intended task to perform, giving the limited capacity to codify the complete reality of the represented knowledge. Therefore, the representation of the knowledge should be limited and defined according to convenience to the given task. In the definition of the representation, the level of formalism of the representation may enable the use of the represented knowledge in different ways (Guarino, 1995). In this study, the functional approach to representation was adopted to enable further analysis of the represented knowledge (Hayashi & Ohsawa, 2015), as defined in later section.

2.2.Named Entity Recognition

In Natural Language Processing (NLP), Named Entity recognition and Classification (NERC) is a kind of task that identifies entities according to predefined classes. The use of textual features support the identification of entities.

Possessive Noun Phrases (PNP) are the expressions for possessive relationships possessor-possessed. With the use of textual features, such as markers and syntax , relationships between elements of the sentence can be identified and named. According to WALS (Nichols & Bickel, 2013), there are 4 main locus of marking in PNPs:

1)Possessor is head-marked,
2)Possessor is dependent-marked,
3)Possessor is double-marked, and
4)Possessor has no marking.

Contact: Danilo Eidy Miura, The University of Tokyo, daneidyucb@gmail.com

Markings in PNP explain the existence of various forms of expression of possessive relationships, given the emphasis on of different elements in the phrase.

Given the nature of variables in data analysis in the Market of Data, possessive expressions including pronouns, which are relevant in narratives, are not relevant to our task. The focus of this study is the identification of variables of entities and events, and not possessions of people.

3. Variable Identification

Our aim is to identify and extract well-formed variable names (WFVN) from the natural language text and discover potential variable labels. Recognizing essential elements of a PNP, and defining their relationship and roles, we formally represent the knowledge of the variable.

In order to understand the WFVN, let's consider the variable as an abstract sense of varieties (attributes, properties, features, qualities, etc.) that needs a paired representative (thing of event) that provides a more concrete sense of what variables may vary to. Between the pair, should exist a possessive relation that places a variable as a qualifier of the concrete sense.

Let PNP be represented by the expression *has* (*x* , *y*) where *x* is the possessor noun and *y* is the possessed noun. For the identification of PNP as a WFVN, the Formal Representation of a Variable should satisfy the following conditions of possessor should be a thing (or event) and possessed should be a variable:

$$\text{has}(x, y) \wedge E(x) \wedge V(y) \rightarrow \text{WFVN}(\text{has}(x, y)) \quad (1)$$

1) has (x , y) : Possessor has possessed.
2) *E (x) : Possessor is a thing or an event.*
3) V (y) : Possessed is a variable.

In possessive nouns phrases, we can identify the noun that represents possessor and the noun that represents possessed. To satisfy the condition 1, the possessive relationship will be identified with three patterns of PNPs, as follows:

a)Pd + of + Pr, (ex. temperature of water)
b)Pr + 's + Pd, (ex. water's temperature)
c)Pd + Pr (ex. water temperature)
Pd: Possessed noun
Pr: Possessor noun

In order to satisfy the conditions 2 and 3, it is needed to take in account the relation between the noun and their relative abstractness and concreteness. The relation between Pr and Pd should make a clear distinction between their senses and provide both abstract and concrete sense. This consideration regards the fact of PNPs without clear distinction may not a full sense of the variable:

A) Level of temperature (double abstract)
B) Pool water (double concrete)

In the example A, the level of abstraction of both nouns are high, and we don't have a full sense of the variable, missing the concrete sense of and event (ice, melting, boiling, condensing, …)

In the example B, the problem is on the lack of abstraction. Since both nouns provide concrete sense, we don't have a clear idea possession. This example allows us to have interpretation without possession:
Instead of *Water of pool*,
Interpret W*ater in pool*

4. Experiment on Data Jacket Store

An experiment was designed to verify the performance of the identification of WFVN. Possessive relationships can be identified with use of the three patterns of PNPs, defined before. And regarding the conditions 2 and 3 discussed before, we considered the use of Wordnet hypernyms (Scott & Matin, 1998) as features to establish the concrete-abstract relationship between the nouns in the phrase.

Using a database containing natural language contents and variable names, we could attribute a score to the candidates of variables. Data Jacket Store is a catalog of more than 1000 Data Jackets (Hayashi & Ohsawa, 2015), digest information about the datasets that contain natural language description of the data, as well as variable labels.

In the experiment, we tested the use of hypernyms as features to distinguish PNPs that represent WFVN from the others. Assuming features of WFVN supports the identification of new variable names with similar features, we defined the probability of a new PNP be a WFVN is defined by the probability of the new PNP to have similar hypernyms of WFVN.

Given *n* as a noun in a new PNP, H as a *i* number of hypernyms of *n*, extracted from Wordnet, *v* is the condition of being a variable, and *e* is the condition of being a thing (or event). We define the probability of role (*v* or *e*) of the noun in a new PNP as the average of the probabilities of each PNP's hypernym to be a variable's hypernym:

$$P(v|n) = \sum_{1}^{i} P(v|Hi) * P(Hi|n) \qquad (2)$$

$$P(e|n) = \sum_{1}^{i} P(e|Hi) * P(Hi|n) \qquad (3)$$

The probability of a given hypernym to be a variable's hypernym can be defined by the probability of the given hypernym be the VLs' hypernym.

4.1. Procedures

The experiment was design to demonstrate the performance of the Classifier with the use of trained data.

1) Define the probability of hypernyms to be variable's hypernyms by the distribution of hypernyms of VLs of DJ Store.

2) Identify and extract PNPs in the DJ's outlines.

3) Calculate the probability of the PNPs to be variables, according to the equation in previous session.

4) Classify PNPs according to the existence in VLs list and satisfaction of the following criteria:
   $P (v \mid n) > P (\neg v \mid n)$

5. Results of DJ Store Experiment

In total of 1032 DJ Outlines, 8660 PNPs were identified according to the three patterns of PNP. In total, DJ Store provided 5836 Variable Labels from which 3788 different representations were extracted.

The formal representation of the variable names solved the problem of variances in the expression of possessive relationships due to the markings of the language (Nichols & Bickel, 2013). The formal representation eliminates markings and different patterns of PNP are represented in the same way. Possessive Noun Phrases such as *temperature of water*, *water's temperature* and *water temperature* will be uniquely represented as has (water , temperature).

Regarding the performance of the experimental test to discover potential Variable Labels, discovery is shown as the identification of variables that does not exist in DJ Store VL list. Using the formal representation of variable, we could discover 2017 new PNPs that satisfied the defined criteria. It suggests the information from DJ Outlines shows latent Variable Labels that can be considered in the utilization of that data.

Examples of new identified variables:
Temperature of air,
Efficiency of fuel,
Number of climbers,
Number of surgeries, …

6. Discussion and Future Work

In this paper, the proposal of use of Possessive Relationships as feature of variable names was demonstrated through a simple experiment using DJ Store. And results point to a potential use of the possessive relations as feature to identify variables in text. But it suggests the need of improvements in the selection of candidate relations.

In future work, we aim to refine variable selection with better understanding of types of variables and more accurate algorithms. We also should consider variables that are not considered measurable, such as classes or unstructured data.

Acknowledgements

This work was made possible thanks to to the initiatives of Data Jacket Promotion Work Group which provided the access to the DJ Store data for experiments and it was funded by JST CREST No. JPMJCR1304, JSPS KAKENHI JP16H01836, and JP16K12428, and industrial collaborators.

References

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. International journal of human-computer studies, 43(5-6), 625-640.

Johanna Nichols, Balthasar Bickel. 2013. Locus of Marking in Possessive Noun Phrases. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/24, Accessed on 2019-01-28.)

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. Data and knowledge engineering, 25(1), 161-198.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation?. AI magazine, 14(1), 17.

Hayashi, T., & Ohsawa, Y. (2015). Knowledge Structuring and Reuse System Using RDF for Supporting Scenario Generation. Procedia Computer Science, 60, 1281-1288.

Scott, S., & Matwin, S. (1998). Text classification using WordNet hypernyms. Usage of WordNet in Natural Language Processing Systems.

Using Sequence Constraints for Modelling Network Interactions

Johannes De Smedt[*1] Junichiro Mori[*2] Masanao Ochi[*2]

[*1] The University of Edinburgh [*2] The University of Tokyo

The ubiquitous nature of networks has led a vast number of works dedicated to the study of capturing their information. Various graph-based techniques exist that report on the characteristics of nodes and edges, e.g., author-citation networks, social interactions, and so on. A significant amount of information can be extracted by summarizing the surrounding network structure of nodes, e.g., by capturing motives, or walk patterns. In this work, we present a new way of capturing the interaction between nodes in a network by making use of the sequence in which they occur. (1) The objective of this paper is to make use of behavioural constraint patterns; a concise but detailed report of node's interactions can be constructed that can be used for various purposes. (2) It is shown how the constraint patterns can be mined form interaction data, and how they can be used for various applications.

1. Introduction

Networks are often formed by the interaction of various actors. For example, social networks grow based on friendship or interested-based relations, forum posts and emails link users according to their communication patterns, and citation networks are formed through authors referencing peers in their field. Typically, the construction of these networks is based on either undirected, or directed edges with weights. Furthermore, many network techniques focus on static relationships, I.e., the evolution over time is not investigated. However, a range of new techniques emerged recently that focus on the time-aspect of a network. Most notably, the use of motifs [Paranjape 17], and streams [Latapy 18] allow to capture the evolution of a network over time. In this paper, we describe a new approach based on behavioral constraints, I.e., constraints based on sequence patterns that allow to describe the order of the interactions of nodes.

We investigate how they can be constructed from a network dataset, and use the various patterns to describe the evolution of the network over time. In particular, we apply the sequence mining method to the question-and-answer interaction-based network. Our preliminary results show that profiling network interactions patterns with sequence mining enables track the behaviour of nodes in a transactional network without relying on the typical partial-order based results.

This paper is structured as follows. In Section 2, the methodology is presented to mine constraints from network data. Next, Section 3 reports on the application on real-life datasets. Section 4 concludes the paper and reports on the future directions.

2. Behavioural constraint patterns in networks

In this section, a detailed overview of the constraints is given, and how they can be leveraged for various network analysis applications.

2.1 Constraint set

Behavioural constraint templates have been long used in various areas of computer science. Most notably, a comprehensive set of Linear Temporal Logic (LTL) templates was proposed for the formal verification of program execution [Dwyer 99]. LTL provides an adequate formalism to search for various temporal properties, such as whether something happens eventually, next, and so on, and can be used in conjunction with typical logical operators to construct expressive relations. The initial set was extended to include various other relations, most notably unary ones. While initially proposed as LTL formulae which are convertible to Büchi automata, finite trace equivalent regular expressions were introduced in [Di Ciccio 13]. Models allowing for multiple constraints at the same time can be obtained by conjoining the automata to obtain a global language or automaton, over which all constraints hold.

In Table 1, an overview of the most-commonly used constraints in literature. They are organized according to 7 different categories, including unary and binary constraints. Most notably, the binary constraints are exhibit a hierarchy which is reported in [Di Ciccio 13] and which covers unordered up to chain ordered (using the next operator). Besides, the inclusion of negative constraints is unique, as typically only existing patterns are reported. Including negative behaviour can be used to find relations that are not apparent at first sight, e.g., in Figure 1 , the fact that nodes A and E are both present in the sequence of C, but do not have interactions themselves, still allows the inference of not succession(A,E).

Despite not being useful for capturing interaction effects, the unary constraints can be used for adding information to a node's feature vector in case any exist. I.e., if a particular node is always occurring first in a sequence, this might signify a particular pattern, e.g., a person reporting recently-occurred disasters.

Not every constraint is suitable for binary interaction within a network context, i.e., not chain succession is, in general, not suitable for profiling behavior, as it holds in many situations. Besides, absence is hard to identify unless a particular node is scrutinized for this behaviour in the sequence of another node. Exclusive choice and not co-

Contact: Johannes De Smedt, The University of Edinburgh, johannes.desmedt@ed.ac.uk

Table 1: An overview of Declare constraint templates with their corresponding regular expression.

Template	Regular Expression
Existence(A,n)	.*(A.*){n}
Absence(A,n)	[ˆA]*(A?[ˆA]*){n-1}
Exactly(A,n)	[ˆA]*(A[ˆA]*){n}
Init(A)	(A.*)?
Last(A)	.*A
Responded existence(A,B)	[ˆA]*((A.*B.*) |(B.*A.*))?
Co-existence(A,B)	[ˆAB]*((A.*B.*) |(B.*A.*))?
Response(A,B)	[ˆA]*(A.*B)*[ˆA]*
Precedence(A,B)	[ˆB]*(A.*B)*[ˆB]*
Succession(A,B)	[ˆAB]*(A.*B)*[ˆAB]*
Alternate response(A,B)	[ˆA]*(A[ˆA]*B[ˆA]*)*
Alternate precedence(A,B)	[ˆB]*(A[ˆB]*B[ˆB]*)*
Alternate succession(A,B)	[ˆAB]*(A[ˆAB]*B[ˆAB]*)*
Chain response(A,B)	[ˆA]*(AB[ˆA]*)*
Chain precedence(A,B)	[ˆB]*(AB[ˆB]*)*
Chain succession(A,B)	[ˆAB]*(AB[ˆAB]*)*
Not co-existence(A,B)	[ˆAB]*((A[ˆB]*) |(B[ˆA]*))?
Not succession(A,B)	[ˆA]*(A[ˆB]*)*
Not chain succession(A,B)	[ˆA]*(A+[ˆAB][ˆA]*)*A*
Choice(A,B)	.*[AB].*
Exclusive choice(A,B)	([ˆB]*A[ˆB]*) |.*[AB].*([ˆA]*B[ˆA]*)

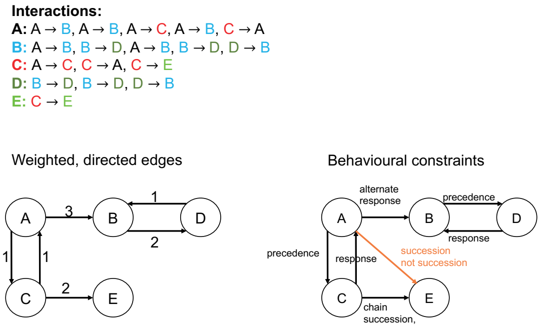


Figure 1: Running example

existence are similar in this respect, where the latter does not require the presence of either. Similar to not chain succession, this might lead to the discovery of many frequently non-occurring pairs.

2.2 Mining the patterns

We define transactional network data as an ordered set of interactions T between nodes from the set N, where each transaction is a tuple $(n_1, n_2, ts) \in T$ with $n_1 \in N$ the initiating node, $n_2 \in N$ the receiving node, and $ts \in \mathbb{N}^+$ a timestamp. T can be read sequentially, where each node $n \in N$ has a sequence $s_n \subset 2^{|N|}$ that is extended whenever a transaction $t \in T$ is for that node is witnessed. I.e., s_n gets extended with $\langle n, n_o \rangle$ whenever n is the initiating node, and with $\langle n_o, n \rangle$ when n is on the receiving end given another node $n_o \in N$.

By using the interesting Behavioural Constraint Miner [De Smedt 17], we can mine all patterns in a a sequence s_n to obtain a set of constraints C_{s_n}. Note, however, that if a given binary constraint $c(n, n_2) \in C_{s_n}$ holds for n in its own sequence, this still has to be verified with the sequence of the other node. If $c(n, n_2)$ is not present in that sequence, the constraints does not hold. Consider for example the interaction in Figure 1. Despite the evidence in the sequence of A that there exists an alternate succession relationship between A and B due to the alternating ABABAAB pattern, the sequence of B rather indicates that other occurrences of B happen in between (e.g. B→D), breaking the pattern. Hence, a final step is required to recursively ensure that $C_n = \{c \mid c \in C_n \wedge c \in C_{n_i} \forall n_i \in \mathcal{N}(n) \vee c \notin C_n \wedge c \notin C_{n_i} \forall n_i \in \mathcal{N}(n)\}$ where $\mathcal{N}(n) \subseteq N$ denotes the neighbourhood of node n to check that all constraint pertaining to n are either both in its constraint set and the constraint set of its neighbours to avoid conflict, or that it is present in an unrelated node (e.g. the connection succession(A,E) in Figure 1). To conclude the discovery of sequence templates from the network interactions, the sets C_n are pruned according to the constraint hierarchy.

2.3 Applications

The mining of interactions in a network as sequences has several applications. Most notably, the sequence information can be used for analyzing the interactions' evolution over time. By tracking what patterns exist, and whether they return over time gives an overview of how certain relations change and what the underlying sequential behaviour is.

Next, the sequence patterns can be used as features of a node. In this case, also unary constraints help define the node in terms of where in a sequence, how often, and with what other nodes the node is interacting. The features can be used towards node classification [Bhagat 11]. Finally, by using the transitivity properties of the constraints, link inference/prediction [Liben 07] can also be made.

3. Results

We apply the sequence method to the Math Overflow dataset, as used in [Paranjape 17]. On the Overflow web sites, users post questions and receive answers from other users, and users may comment on both questions and answers. We derive a transactional network by creating an edge (u, v, t) if, at time t, user u: (1) posts an answer to user vs question, (2) comments on user vs question, or (3) comments on user vs answer. The data contains 24,818 nodes with 506,550 interactions over 2,350 days and deals with question-and-answer data from users regarding mathematical problems.

We retrieve the constraints over the dataset by splitting the interactions into contingent blocks of a varying time length. In this case, we used blocks of 4 hours, 2 days, 100 days, and 1,000 days in order to track the evolution of the constraints. For this analysis, we limit the constraint set to the 7 most common sequence patterns. In order to illustrate the usefulness of the results, we focus on two active users with a different background. The first user (denoted B) is considered an authority as that node in the network has the highest authority score [Ding 04]. The high authority is pointed to by many high hubs and high hub points to many high authorities. Authority and hub scores are obtained by this iterative scoring.

Table 2: An overview of the proportion of constraints that shift from one sequence pattern into another, both for incoming and outgoing constraints of nodes A and B. The colours denote the place in the distribution, where red is higher and green lower. Scores with different colours and equal scores indicate a difference in value behind the significant digits.

	In								Out							
4 hours - A	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
NotSuc (1)	0.15	0.04	0.01	0.00	0.02	0.02	0.00	0.01	0.14	0.05	0.05	0.00	0.02	0.02	0.00	0.04
Prec (2)	0.14	0.03	0.02	0.00	0.01	0.01	0.00	0.01	0.13	0.01	0.03	0.00	0.01	0.02	0.00	0.02
AltPrec (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
ChainPrec (4)	0.03	0.03	0.02	0.00	0.03	0.02	0.00	0.00	0.09	0.02	0.05	0.00	0.02	0.03	0.00	0.04
Resp (5)	0.15	0.02	0.01	0.00	0.01	0.03	0.00	0.01	0.11	0.01	0.00	0.00	0.00	0.03	0.00	0.00
AltRes (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes (7)	0.04	0.05	0.01	0.00	0.01	0.02	0.00	0.04	0.13	0.00	0.02	0.00	0.01	0.02	0.00	0.15
4 hours - B																
NotSuc	0.17	0.02	0.02	0.00	0.01	0.01	0.00	0.01	0.21	0.02	0.03	0.00	0.01	0.02	0.00	0.01
Prec	0.22	0.02	0.02	0.00	0.01	0.01	0.00	0.01	0.19	0.01	0.02	0.00	0.00	0.01	0.00	0.01
AltPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.05	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.06	0.01	0.01	0.00	0.00	0.01	0.00	0.01
Resp	0.21	0.02	0.01	0.00	0.00	0.02	0.00	0.01	0.20	0.02	0.01	0.00	0.00	0.02	0.00	0.00
AltRes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.07	0.01	0.01	0.00	0.00	0.01	0.00	0.03	0.06	0.01	0.01	0.00	0.00	0.02	0.00	0.02
2 days - A																
NotSuc	0.19	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.26	0.01	0.01	0.00	0.00	0.01	0.00	0.00
Prec	0.32	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.24	0.01	0.01	0.00	0.00	0.01	0.00	0.00
AltPrec	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.33	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.25	0.01	0.00	0.00	0.00	0.02	0.00	0.00
AltRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2 days - B																
NotSuc	0.21	0.02	0.02	0.00	0.00	0.02	0.00	0.00	0.23	0.01	0.02	0.00	0.00	0.02	0.00	0.00
Prec	0.29	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.27	0.02	0.02	0.00	0.00	0.01	0.00	0.00
AltPrec	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.29	0.02	0.01	0.00	0.00	0.02	0.00	0.00	0.28	0.02	0.00	0.00	0.00	0.03	0.00	0.00
AltRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Similar overview as Table 3 containing the 100 and 1,000 days time frames.

	In								Out							
100 days - A	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
NotSuc	0.13	0.02	0.03	0.00	0.00	0.04	0.00	0.00	0.18	0.04	0.04	0.00	0.00	0.04	0.00	0.00
Prec	0.24	0.04	0.07	0.00	0.00	0.02	0.00	0.00	0.19	0.04	0.05	0.00	0.00	0.01	0.00	0.00
AltPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.25	0.04	0.02	0.00	0.00	0.07	0.00	0.00	0.18	0.04	0.01	0.00	0.00	0.04	0.00	0.00
AltRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100 days - B																
NotSuc	0.10	0.02	0.04	0.00	0.00	0.05	0.00	0.00	0.15	0.05	0.06	0.00	0.00	0.05	0.00	0.00
Prec	0.19	0.04	0.08	0.00	0.00	0.03	0.00	0.00	0.16	0.05	0.07	0.00	0.00	0.02	0.00	0.00
AltPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.22	0.05	0.02	0.00	0.00	0.11	0.00	0.00	0.13	0.04	0.02	0.00	0.00	0.05	0.00	0.00
AltRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000 days - A																
NotSuc	0.07	0.03	0.06	0.00	0.00	0.07	0.00	0.00	0.11	0.04	0.07	0.00	0.00	0.06	0.00	0.00
Prec	0.14	0.04	0.09	0.00	0.00	0.06	0.00	0.00	0.12	0.06	0.09	0.00	0.00	0.03	0.00	0.00
AltPrec	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.17	0.06	0.04	0.00	0.00	0.14	0.00	0.00	0.10	0.05	0.04	0.00	0.00	0.07	0.00	0.00
AltRes	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000 days - B																
NotSuc	0.01	0.03	0.01	0.00	0.00	0.05	0.00	0.00	0.06	0.02	0.19	0.01	0.00	0.09	0.00	0.00
Prec	0.08	0.06	0.05	0.00	0.00	0.12	0.00	0.00	0.05	0.02	0.23	0.01	0.00	0.03	0.00	0.00
AltPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainPrec	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Resp	0.14	0.10	0.04	0.00	0.00	0.29	0.01	0.00	0.02	0.01	0.05	0.00	0.00	0.05	0.00	0.00
AltRes	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ChainRes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The other user (denoted A) has a similarly high degree (high number of connections in the network), but a lower authority score. Our hypothesis is that the interactions of the authority user result in several constraint patterns as he gains the authority through answering and commenting to questions within his expertise.

The results of the shifts in constraint patterns as expressed in their proportions, are included in Tables 2 and 3 for both incoming and outgoing constraints of both nodes. The cells indicate the proportion of connections between the same nodes that are both present again in two subsequent time frames that shifted from the template in the rows, to the template in the columns. '0' signifies that the constraint is no longer present between both rows.

Firstly, it can be seen that there is a high number of constraints that are not reoccurring over time, meaning they are not repeated in the subsequent time frame. This behaviour is expected, given that many question-answering threads stop after a few posts, and many users only tend to intervene in a limited number of threads. Considering different lengths of time frames, however, we note that especially for node B (the authority) the number of vanishing interactions is drastically lower for 1,000 days. In case of incoming constraints, we see many re-occurring response constraints, and with outgoing ones we see many precedence and not succession constraints appearing. This is in line with how we would expect question-answering is handled by an authority, who responds to all questions within his area of expertise.

Overall, the two nodes behave relatively similarly in terms of proportions of constraints up until the 1,000 days threshold. The change incurred by increasing the time frames does not yield drastically different results, but it can be noted that more connections are reoccurring (mostly response and precedence relationships) rather than vanishing (as captured by column '0'). Hence, nodes that are surviving longer, and hence are reoccurring themselves, seem to maintain their relations over time. Also, any 'stronger' constraints that model alternating or chain relations are very often not present. One final observation is interesting. The high number of chain response connections that are going out from node A indicates that many immediate answer-response messages were exchanged over a period of 4 hours, indicating that single conversations where picked up of which many reoccurred as well.

4. Conclusion and future work

In this paper, we have shown how mining network interaction patterns can be profiled using sequence mining techniques. We apply the sequence mining method to the question-and-answer interaction-based network. Our preliminary results show that employing sequence patterns enables us track the behaviour of nodes in a transactional network and summarize their interactions without relying on the typical partial-order based results that are offered in sequence mining, while still going beyond the typical general nature of motifs that focus on directed arcs between 2 or 3 actors [Paranjape 17] . In a small experimental evaluation, we demonstrate the usefulness of the approach in the context of message board analysis.

For future work, we envision to focus on testing the patterns in the context of feature engineering, and link inference.

References

[Paranjape 17] Paranjape, A., Benson, A. R., & Leskovec, J.: Motifs in temporal networks, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (2017)

[Namaki 17] Namaki, M. H., Wu, Y., Song, Q., Lin, P., & e, T.: Discovering graph temporal association rules, Proceedings of the 2017 ACM Conference on Information and Knowledge Management (2017)

[Dwyer 99] Dwyer, M. B., Avrunin, G. S., & Corbett, J. C.: Patterns in property specifications for finite-state verification, Proceedings of the 21st international conference on Software engineering (1999)

[Latapy 18] Latapy, M., Viard, T., & Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. Social Network Analysis and Mining, 8(1) (2018)

[Di Ciccio 13] Di Ciccio, C., & Mecella, M.: A two-step fast algorithm for the automated discovery of declarative workflows, 2013 IEEE Symposium on Computational Intelligence and Data Mining (2013)

[De Smedt 17] De Smedt, J., Deeva, G., & De Weerdt, J.: Behavioral Constraint Template-Based Sequence Classification, European Conference on Machine Learning (2017)

[Ding 04] Ding, C. H., Zha, H., He, X., Husbands, P., & Simon, H. D.: Link analysis: hubs and authorities on the World Wide Web, SIAM review, 46(2) (2004)

[Bhagat 11] Bhagat, S., Cormode, G., & Muthukrishnan, S.: Node classification in social networks, Social network data analytics, Springer (2011)

[Liben 07] LibenNowell, D., & Kleinberg, J.: The linkprediction problem for social networks, Journal of the American society for information science and technology, 58(7) (2007)

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

CTransE : Confidence-Based Translation Model for Uncertain Knowledge Graph Embedding

Natthawut Kertkeidkachorn[*1*2] Xin Liu[*1] Ryutaro Ichise[*2*1]

[*1] National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
[*2] National Institute of Informatics, Tokyo, Japan

Knowledge graphs play an important role in many AI applications such as fact checking. Many studies focused on learning representations of a knowledge graph in a low-dimensional continuous vector space. However, most of the recent studies do not learn embedding representations on uncertain knowledge graphs. Uncertain knowledge graphs, e.g., NELL and Knowledge Vault, are valuable because they can automatically populate themselves with new facts. Nevertheless, the automatic process basically induces uncertainty to knowledge. In this study, we introduced knowledge graph embedding on uncertain knowledge graphs by using adapting confidence-margin-based loss function for translation-based models, namely CTransE, to deal with uncertainty on knowledge graphs. The results show that CTransE can robustly learn representations of uncertain knowledge graphs and outperforms the conventional method on knowledge graph completion task.

1. Introduction

A knowledge graph is a structured knowledge base, which provides real-world facts as knowledge. A knowledge in a knowledge graph is represented as a triple (h, r, t), where h and t are entities and r is a relation directed from h to t. Such knowledge has been widely used in many recent AI applications such as fact checking. Since Knowledge graphs become popular, the research community has made a great effort in constructing them. Currently, there are many publicly available knowledge graphs, such as DBpedia and Freebase. However, these knowledge graphs require manual effort to curate and keep up-to-date.

Unlike the above efforts, other approaches try to automatically build knowledge graphs [3]. However, automated construction of Knowledge Graphs often results in noisy and inaccurate facts, whose degree of reliability can be expressed by a score. The well-known uncertain knowledge graphs are Reverb and NELL.

Recently, knowledge graph embedding has gained the attention of many researchers. Knowledge graph embedding learns to capture latent representations of triples in a knowledge graph by projecting the entities and the relations of triples in the knowledge graph to a continuous low-dimensional vector space without considering the uncertainty of a knowledge graph. Generally, the uncertainty of a triple provides the reliability of the triple. Ignoring such reliability, noisy triples could induce the problem on the representation learning process.

In this paper, we introduce a confidence margin-based loss function on the translation model, namely CTransE, to deal with the uncertainty of triples in Knowledge Graphs. In CTransE, an uncertainty score is treated as the weight for a triple. The higher weight is, the lower the uncertainty is. A higher weight means that it is more likely a triple is true. CTransE handles the weight by adjusting the margin of the translation model in order to encode uncertainty into the representation.

2. Problem Definition

Given an uncertain knowledge graph denoted by $G = (E, R, Q)$, where E, R, and Q are the entity set, relationship set, and fact set, respectively. A fact is represented by a quadruple $q = (h, r, t, s)$, where $h, t \in E$, $r \in R$, and $s \in \mathbb{R}_{[0,1]}$. It indicates that entities h and t are connected by a relation r with score s, uncertain knowledge graph embedding is to learn embedding representations of an entity $\vec{e} \in \mathbb{R}^K$ for each $e \in E$ and a relation $\vec{r} \in \mathbb{R}^K$ for each $r \in R$ such that for each $(h, r, t, s) \in Q$; $f(h, r, t) \propto 1 - s$, where $f(h, r, t)$ is any arbitrary score function for q, such as $|\vec{h} + \vec{r} - \vec{t}|$, i.e. the facts can be preserved in $\mathbb{R}^K$ while considering their confidence.

3. Related Work

One of the popular models for knowledge graph embedding is the translation model. The translation models embed representations by using the relation r from the head entity h to the tail entity t as a dissimilarity score. The first model for the translation model is TransE [1]. TransE computes the triple's dissimilarity score by (h, r, t) as $\vec{h} + \vec{r} = \vec{t}$. With this translation, it can capture the first-order rules. Later, there are many models improving TransE by proposed the different dissimilarity functions.

However, such models are not supported uncertain knowledge graphs. In an uncertain knowledge graph, the level of reliability of a fact is represented in terms of a confidence s. So far, the translation methods do not take the confidence s of each fact into account. In practice, we can ignore the confidence of the facts and learn the embedding. Nevertheless, without confidence as an indicator, noisy facts can degrade the quality of the embedding representations. In this study, we therefore aim to introduce a new margin-

Contact: Natthawut Kertkeidkachorn, natthawut@nii.ac.jp

based loss function for supporting the uncertain knowledge graph embedding on the translation models.

4. Uncertain Knowledge Graph Embedding

The confidence margin-based translation model (CTransE) is to improve the margin-based loss function in translation models in order to support confidence on the quadruple q. The margin-based loss function is as follows.

$$L=\sum_{(h,r,t)\in T}\sum_{(h',r,t')\in T'}[f(h,r,t)-f(h',r,t')+M]_{+} \quad (1)$$

, where $[x]_{+}$ is the positive part of x, $f(\cdot)$ is a score function, M is a margin, and (h',r,t') is a negative sample in T'. To preserve the embedding in the vector space, TransE uses normalization as the regularization in each iteration.

As shown in Eq. 1, the margin-based loss function does not consider the score s in the quadruple q. As a result, the reliability of triples is ignored. To overcome this problem, we propose a confidence margin-based loss function for translation models by varying the margin M of each triple based upon the score s. The idea behind is that the higher the uncertainty of the quadruple, the less margin should be used to keep the relation because (e,e') is likely to be noise. The relation r then should not be held with the margin M due to such uncertainty. We therefore derive the confidence margin-based loss function as follows.

$$L=\sum_{(h,r,t,s)\in Q}\sum_{(h',r,t',s)\in Q'}[f(h,r,t)-f(h',r,t')+sM]_{+} \quad (2)$$

where $[x]_{+}$ is the positive part of x, $f(\cdot)$ is the score function for (h,r,t) of the quadruple q, M is the margin, $(h',r,t',1.0)$ is a negative sample in Q' generated in the same way as T' and s is the confidence of the quadruple.

5. Experiments and Results

To evaluate CTransE for learning embedding representations for an uncertain knowledge graph, we conducted the experiment knowledge graph completion. Knowledge graph completion is a task to fill the knowledge graph by predicting missing relationships between entities. Given an incomplete uncertain knowledge graph G, the task is to fill in G by predicting the set of missing quadruples $Q'=\{(h,r,t,\cdot)\mid h,t\in E, r\in R,(h,r,t,\cdot)\notin Q\}$.

Currently, there are many datasets for the knowledge graph completion. However, these datasets do not contain uncertainty of triples. We, therefore, constructed the dataset from a real knowledge graph, NELL [2]. NELL provides a confidence score for each triple. To build our dataset, we first collected quadruples form NELL at the 995^{th} iteration. Then, we followed the cleaning process [4]. However, we did not add the inverse relation to the dataset as was done in that study. As a result, we obtained 75,491 entities, 200 relations, 134,213 training, 10,000 validation, and 10,000 testing quadruples.

Table 1: Results of knowledge graph completion

Method	% Hit@		MR
	1	10	
TransE	10.44	30.15	0.175
CTransE	**11.11**	**30.47**	**0.180**

The experimental setup and the evaluation protocol of the experiment are similar to the study in TransE [1]. Although our confidence-margin-based loss function can be applied to any arbitrary translation models, we select TransE to study due to its simplicity. As a result, the dissimilarity function in the experiment is set as L1-norm and TransE becomes the baseline for the experiment. The implementations of TransE and CTransE both used the grid search algorithm to find appropriate parameters. The dimension was selected from {20,50,100,200}. The search range for the margin M was set at {1,5,10,50,100}. The learning rate was selected from {0.1,0.001,0.0001}. In the evaluation process, we employed three evaluation metrics: Hit@1, Hit@10 and mean reciprocal rank (MR) as the study [1].

The experimental result is presented in Table 1. The result shows that CTransE outperforms TransE. This result indicates that the confidence of the triples affects the learning representation on uncertain knowledge graph and CTransE can capture such uncertainty to improve embedding representations.

6. Conclusion

We introduced a new confidence-margin-based loss function, namely CTransE, for the translation model. The preliminary results show that CTransE could encode the uncertainty of knowledge graphs and that better learn the embedding representation than the traditional margin-based loss function on uncertain knowledge graph.

References

[1] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[2] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, pages 1306–1313, 2010.

[3] N. Kertkeidkachorn and R. Ichise. An automatic knowledge graph creation framework from natural language text. *IEICE Transaction on Information and Systems*, 101(1):90–98, 2018.

[4] W. Xiong, T. Hoang, and W. Y. Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *EMNLP*, pages 564–573, 2017.