Reducing Miscommunication in Cross-Disciplinary Concept Discovery using Network Text Analysis and Semantic Embedding

Akkharawoot Takhom[1,2], Prachya Boonkwan[3], Mitsuru Ikeda[1], Sasiporn Usanavasin[2], and Thepchai Supnithi[3]

[1] School of Knowledge Science
Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan
{akkharawoot, ikeda}@jaist.ac.jp
[2] School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand
sasiporn.us@siit.tu.ac.th
[3] Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thai-land
{prachya.boonkwan, thepchai.supnithi}@nectec.or.th

Abstract. Multiple thinking of different stakeholders has to influence collaborative working. Some part of information exchange is fragment knowledge that is a significant challenge to complete knowledge co-creation from various domains. However, an effective obstacle is miscommunication among the stakeholders, particularly when ambiguous terms are mentioned in the discussion contexts. To overcome the challenge, this paper proposes an integration approach of network text analysis and knowledge graph embedding. The approach is employed for understanding semantic meaning of terms from a source of knowledge, as a discussion forum. We calculate each term detected by our cross-domain codebook onto the vector space and straightforwardly investigate the relationship among questions and answers on it. To demonstrate the benefits of employing the approach, the system's functionality is implemented to manifest a capability of detecting and reducing miscommunication by the case study of Life Cycle Assessment's discussion board.
Keywords: multidisciplinary knowledge, concepts extraction, Semantic Embedding, cross-disciplinary concept, sustainable development, natural language processing

1 Introduction

An understanding of multiple disciplinary perspectives has influenced in stakeholder success and research goals achievement. Different stakeholder shared their knowledge during discussion based on their expertise, as drawing knowledge across different disciplines. A communication of different stakeholders would be difficult when information is not correct with terms from different disciplines. To understand a problem, a term of *multidisciplinary knowledge* [1] is taken into account in a miscommunication from multiple perspectives of stakeholders, as *a blind spot*, because the knowledge has a limitation within domain boundaries. For example, sustainability science [1] has multiple disciplines, such as environmental protection, economic growth, and human development. To employ the sustainability science, it would be difficult in verifying an understanding of multiple perspectives [2].

For discovering the blind spot, a source of information as *a discussion forum* [3] is a medium allowing domain stakeholders for information exchange and knowledge sharing. Participators can inquire information from other, such as a domain expert. Although a discussion forum is useful for knowledge acquisition, answering contexts can lead them to misunderstand. A *Network Text Analysis (NTA) method* [4, 5] is a method of text mining for discovering a cause of the blind spot from textual data, as follows. First, Aviv et al. [6] apply the method for tracking an interrelation among terminologies in an academic domain. Next, Hecking et al. [7] explore types of users in a discussion forum and analyze them by considering in a network text visualization. Then, Daems et al. [5] use the method to analyze contents to check understanding of science learners.

Although NTA can be exploited in content analysis, there is one of research challenge is to verify the semantic meaning. Several research questions are as follows: (1) how to discover a cause of misunderstanding in discussion contexts, and (2) how to identify terms in a context, which contains multiple domains. In this paper, we are improving NTA by embedding knowledge graph technique to identify a position of semantic concepts in the vector space for considering multidisciplinarity-oriented misunderstandings. An algorithm for knowledge graph completion is *TransR* algorithm [8] that compute the embedding of each semantic concept. Therefore, our research approach attempts to overcome the challenges by proposing an integration approach of NTA and knowledge graph embedding.

The rest of the paper is organized as follows. Section 2 defines multidisciplinary knowledge existing discussion contexts and an analytical approach. Section 3 introduces a cross-disciplinary approach. Section 4 presents a case study in LCA domain for discovering cross-disciplinary concepts. Section 5 discusses an experimental result. Section 6 concludes our research finding and points out future work.

2 Background and Related Works

2.1 Multidisciplinarity in Discussion Contexts

Multidisciplinary knowledge [9] is the knowledge drawing across different disciplines but is limited within domain boundaries. An obstacle of the multidisciplinary knowledge is to understanding a miscommunication from multiple perspectives in the same term of stakeholders in different domains. Regarding the multidisciplinary knowledge, a common term existing in discussion contexts involved two or more disciplines, called *a cross-disciplinary concept* (C^{cd}) [9], has the potential to interlink different domains. A common term exists in discussion contexts involved two or more disciplines.

This paper discovers C^{cd} for analyzing an understanding of various perspectives on participation. As a source of knowledge, *a discussion forum* [3] is an online accessible medium for participants discussion, as information exchange and knowledge sharing. They can inquire questions with and other participants under a topic of interest. For example, a domain expert who has knowledge and experience can answer a relevant question.

However, the obstacle of communication is available when participants discuss in different perspectives. In a situation, all participants discus under the same topic in collaborative work, miscommunication is a cause in collaboration, as *a blind spot.* Therefore, discussion contexts are a crucial source of information in analyzing the problem of miscommunication in different understanding.

2.2 Analytical Approach in a Network Context

Based on a network perspective, we review approaches to textual analysis for representing C^{cd} in textual interrelation. Andresen [5] consider capabilities and accessibility of a discussion forum as follows. First, a huge volume of data is a difficulty for assessment. Second, temporal sequences of the postings, e.g., many answers to one question that a replier may respond to the second answer. Third, time-consuming for information gathering is to measure the quality of a participant's contribution.

Next, a text mining method is an appropriate method contexts analysis from a discussion forum. We are interested in a *network text analysis (NTA)* method [4, 5] for presenting an interrelation among potential terms in domains of interest. Table 1 shown related works by comparing four criteria: (1) interesting domains, (2) using a discussion forum, (3) having the multidisciplinary knowledge, and (4) using NTA. First, Aviv et al. [8] used NTA for data analysis in academic university courses. Next, Chaudhry et al. [10] detect the organizational structure of covert networks. Hecking et al. [7] then explore NTA for analyzing types of users in a discussion forum and visualize the result from collaboratively edited texts. Lastly, Daems et al. [5] use NTA in a contents analysis with domain ontologies for checking an understanding of science learners.

Table 1. A comparison of related works based NTA approach

Related Work	Domain	Discussion Forum	Multidisciplinarity	NTA Method
Andresen [3]	General	X	X	
Hecking et al. [7]	Education	X	X	X
Aviv et al. [6]	Education	X	X	X
Chaudhry et al. [10]	Education		X	
Daems et al. [5]	Education	X	X	X
Our approach	Sustainable Development	X	X	X

In this paper, for determining C^{cd} , we select NTA [5, 7] including natural language processing (NLP) Therefore, NTA method is our appropriate method as contributions in a cross-disciplinary approach and breaking through a blind spot of misunderstanding in multiple perspectives of domain experts.

3 Cross-Disciplinary Approach based on Semantic Embedding

As illustrated in Fig.1, the experimental scenario has an integration of a workflow of NTA and knowledge graph embedding. The workflow has three main parts: NTA phases in orange rectangles, a process of training model in black rectangles, and an experimental result in a green rectangle. The following sections explain in an instruction identified by the ordering numbers in black arrows.

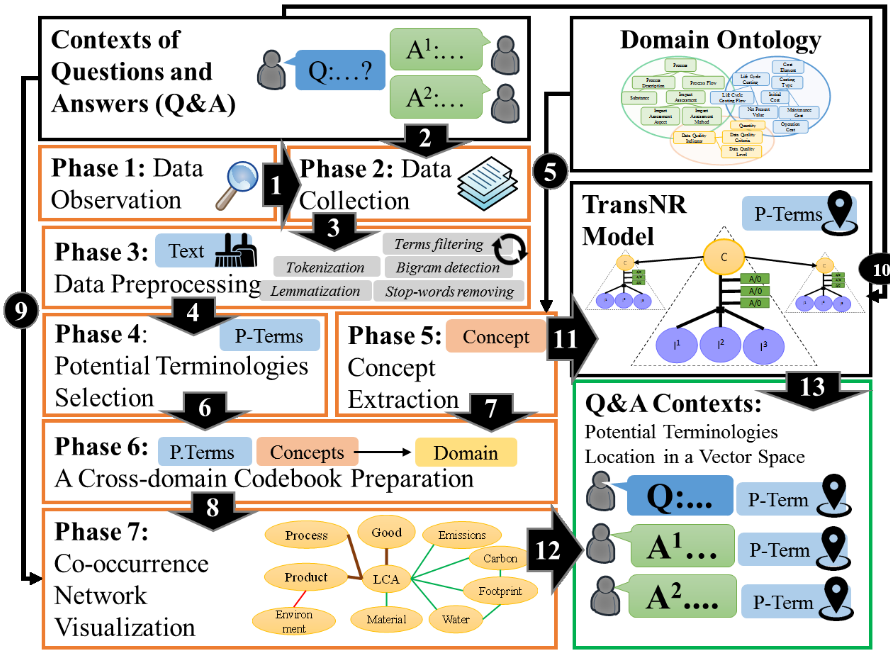

Fig. 1. An overview of the experimental approach: an integration of a workflow of network text analysis [5] and knowledge graph embedding.

3.1 Network Text Analysis

Fig. 1 presents the first part of the experiment that is the NTA workflow in an orange rectangle box. First, *data observation* is Phase 1 to identify sources of knowledge. This phase is surveying tools for data manipulation, such as web crawling or data extraction. Next, Phase 2 is *data collection* that is gathering data from selected sources. A facilitating tool is a scripting language, Python [11], for handling a huge volume of data by using natural language processing. For example, Python's SGML Standard Generalized Markup Language (SGML) parser provides a function for extracting a document markup language. Afterward, Phase 3 is *preprocessing data* in the textual form and has five subphases. A preprocessed result in each phase is explained by the following sentence.

"How to calculate Economic Cost of farming practices during crop production?"

- Subphase 3.1 is *tokenization* that is breaking a stream of text up into meaningful elements or terms, called tokens. The result is "How to calculate Economic Cost of farming practices during crop production?"
- Subphase 3.2 is *lemmatization* that is removing inflectional endings only and then returns the base form of terms, called the lemma. The result is "How to calculate Economic Cost of farming practices during crop production"
- Subphase 3.3 is *stop words removal* that is filtering out unneeded terms. The result is "calculate Economic Cost of farming practices during crop production"
- Subphase 3.4 is *bigram detection* that is detecting a sequence of two adjacent elements from a string of tokens such as two-gram words. Pairs of words are counted by cumulative frequency. The result is "economic cost 31, crop production 25, …, practice crop 2"
- Subphase 3.5 is *terms filtering* that is selecting high-frequency terms from bigram words. The result is "economic cost 31, crop production 25"

Afterwards, Phase 4 is *potential terminology preparation*. A suitable number of high-frequency terms is defined by considering meanings two or more than one domain. The next phase is to consider the meaning of linguistic expressions in natural languages. Regarding a semantic approach [12], an *ontology* is a language to express data fields, concepts, concept relations, and also rules for an inference system allowing us to conduct automated reasoning. To represent concepts in particular meanings, a domain-specific ontology, called *a domain ontology*, represents concepts and semantic relationships between concepts in a semantic network. Therefore, this phase is *concept extraction* (Phase5) exploiting a domain ontology to understand the semantic meaning of a potential terminology. In Phase 6, potential terminologies and extracted concepts are associated for preparing a cross-domain codebook to categorize multiple domains. We can expand relevant terminologies from a domain thesaurus in a case of insufficient concepts. Lastly, Phase 7 is to identify an interrelation of multiple-domains concepts by generating *a co-occurrence network visualization.* To associate the concepts, a number of sliding windows is set and run through the collected contexts. The result is pairs of concepts associated among different domains defined by categories. Therefore, the NTA workflow is the first part of the experiment for analyzing the meaning of C^{cd} in a discussion context. The workflow is straightforward to detect potentially ambiguous terms, which is a cause of misunderstanding.

In the following section, we further detect the domains, which these terms are used ambiguously by projecting a semantic concept into a position in the vector space and measuring similarity.

3.2 Domain Indication with Vector Space Model

Once we detect potentially ambiguous terms in the discussion contexts using the codebook, we now have to determine how misunderstanding takes place during the conversation. We assume that the domain of each text can be indicated by averaging the embedding (i.e., vectors) of each semantic concept occurring in the text. In this paper, we define *miscommunication* as a misunderstanding caused by using terms from other domains with mistaken interpretation.

We directly compute the embedding of each semantic concept via TransR Algorithm [8], an algorithm for knowledge graph completion. In a nutshell, each ontological relation is assigned a separate vector space in which related semantic concepts positon. If two semantic concepts are associated by a semantic relation, they will be projected on the space of such relation and a link between them is established. By means of the vector space model, the relation between two concepts is also represented by a vector which is a subtraction of the destination and source vectors. Symbolically, for any ontological relation r, we project the vectors of two semantic concepts $\mathbf{h}$ and $\mathbf{t}$ to the vector space of r by linear transformation $\mathbf{M}_r$. On each relation r, we attempt to estimate each vectors $\mathbf{h}$ and $\mathbf{t}$ by minimizing the sum of $f_r(\mathbf{h}, \mathbf{t}) = (\| \mathbf{hM}_r + \mathbf{r} - \mathbf{tM}_r \|_2)^2$, where the L_p-norm $\| \mathbf{v} \|_p = (v_1^p + v_2^p + v_3^p + \ldots)^{1/p}$, from the entire knowledge graph. This is an optimization problem and a variety of machine learning techniques have been applied to compute this, e.g. backpropagation and EM Algorithm.

In our method, we integrate all available ontologies for each domain by creating a dummy root node to govern their root nodes. Then we precompute the embedding of each semantic concept via TransR. We will use these vectors to detect misunderstanding in the discussion context.

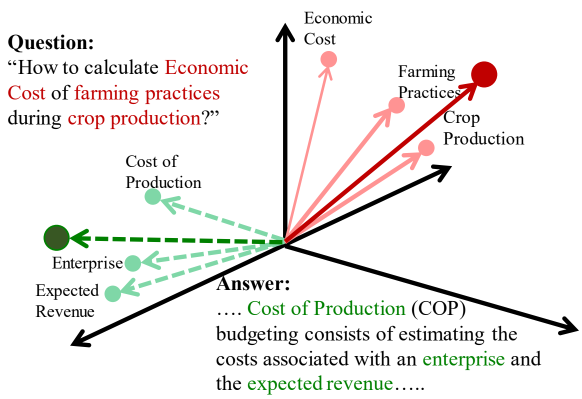

Fig. 2. Vector space model for misunderstanding detection

Next, we will detect each point of miscommunication by averaging the vectors of semantic concepts detected in each text chunk with the codebook. In Fig. 2, suppose there are three terms in the question text "How to calculate *economic cost* of *farming practices* during *crop production*?". Later this question is replied to by the text "The *cost of production* (COP) budgeting consists of estimating the costs associated with an <u>*enterprise*</u> and the <u>*expected revenue*</u>…". Obviously, there is a point of miscommunication caused by ambiguous term "economic cost" that is related to both LCA and economics. To reflect this, we map each term to the corresponding semantic concepts in each text and compute the average vector. As shown, the average vectors of the question and answer texts significantly differ from each other, reflecting cross-domain miscommunication.

4 An Empirical Case Study

4.1 Multidisciplinarity in a Paradigm of Sustainable Development (SD)

An empirical case study is interested in a paradigm of sustainable development (SD) [13] involving in one more than two domains. SD is to preserve environmental resources and to consider human development. Three main aspects includes in SD: economic growth, social development, and environmental protection.

In an environmental protection aspect, *Life Cycle Assessment (LCA)* is a method to quantify energy, a material used, and environmental pollution. Other relevant domains can exploit LCA by following LCA standard guidelines [14].

Although we have the standard guidelines, LCA stakeholders interpret LCA in different perspectives. Miscommunication is a problem of miscommunication when LCA involves in activities of the research or business. With this reason, only a single domain as LCA cannot address a gap of miscommunication. In the following section, we present an experiment scenario from a discussion context of LCA stakeholders.

4.2 An Experimental Scenario

To present multidisciplinarity in LCA, we first observe sources of knowledge (Phase 1) from websites facilitating a discussion forum. We select ReseachGate [15], a social-networking website allowing members (e.g., researchers and scientists) to discuss with each other by posting question or suggestions, as shown in Table 2.

In Phase 2, Python web clawing tool is used for gathering a discussion context, and to extract a data structure in *questions and answers (Q&A)* pages. We collect 148 questions and 92 replies from the Q&A pages under the topic of "Life Cycle Assessment" and "LCA" from September 10, 2016, to October 30, 2016.

Next, we preprocess the collected data (Phase 3), and the results is 22,269 terms by filtering a term-frequency more than 20, as shown in Table 3. Potential terminologies are selected (Phase 4) for LCA and economic domain.

To employ domain ontologies, LCA ontologies are surveyed, and we choose *Data-Qualification for LCA (DQ-LCA)* ontology [16]. The ontology has a characteristic of the multidisciplinary knowledge. As illustrated in Fig. 3, the ontology consists of two domains: LCA domain in a green circle and DQI domain in a yellow group. 396 necessary concepts are extracted (Phase 5) from the ontology by using OWL API [17]. However, extracted concepts are not sufficient for associating an economic domain. We gather economic terminologies from Wikipedia [18] containing 787 economic terminologies and then match them with LCA concepts. After that, a cross-domain codebook is constructed by an association of the potential terminologies and the extracted concepts, and categorized relevant domains for constructing a cross-domain codebook (Phase 6).

Table 2. An example of question answering (Q&A) contexts [15]: economic terms (red italic) and LCA terms (green italic).

Topic:	**Life Cycle Assessment**
Question:	How to calculate *economic cost* of farming practices during *crop production?* I want to calculate economic cost of crop production from soil preparation to crop harvest according to *life cycle assessment (LCA)*, but I am confused as follows: (1) Is there a term "economic footprint" to define this estimation, like carbon footprint?
Answer:	Choosing what *crops* or livestock to produce is an essential decision of any *farm business*. One critical *factor* in making that decision is the *cost of producing* the "enterprises" being considered. This is known as enterprise budgeting or cost of *production budgeting*. Enterprises are a single *crop* or livestock commodity that *produces* a marketable *product*. *Cost of Production (COP)* budgeting consists of estimating the costs associated with an enterprise and the expected revenue. This Factsheet outlines the process and use of *COP budgeting* for farm-level decision-making.

Table 3 A comparison between two domains with term frequency

Pair of Bigram	**LCA**	**Frequency**	**Economic**	**Frequency**
Waste, Management	Management	20	Waste	7
Emissions, LCA	LCA	13	Emissions	24
LCI, Inventory	LCI	17	Inventory	10
Footprint, Product	Product	8	Footprint	37
Waste, Energy	Waste	9	Energy	21
Production, Environmental	Production	9	Environmental	34
Social, Material	Social	4	Material	23
Management, Waste	Management	20	Waste	7

Lastly, we use GePhi [19], a network visualization, to generate a co-occurrence network (Phase 7) by using the cross-domain codebook and the discussion contexts. Fig. 4 represents the result of the experimental scenario in pairs of two relevant domains between LCA and economic domains.

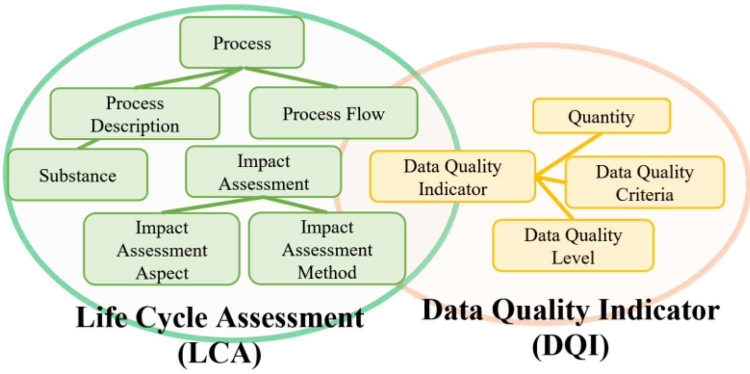

Fig. 3 An excerpt of two upper concepts in the DQLCA ontology [16]: LCA domain in a green circle, and DQI domain in a yellow circle.

The result from NTA is used in the second part. We determine miscommunication during the conversation by computing the embedding of each semantic concept via TransR Algorithm. We integrate DQ-LCA ontology by creating a dummy root node to govern their root nodes, precompute the embedding of each semantic concept via TransR. These vectors are used to detect misunderstanding in the discussion context.

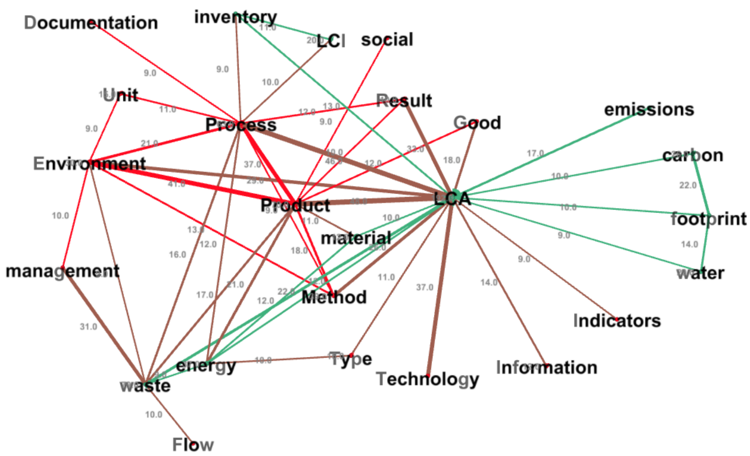

Fig. 4 Co-occurrence network visualization generated by GePhi [19]

5 Results and Discussion

The experimental result presents the co-occurrence network visualization, as illustrated in Fig. 4, that has adjacent edges filtered weights more than 20 and a sliding window with the size of 8 words. These adjacent edges identify an interrelationship between two domains (LCA and economic) in three colors. Corresponding numbers are defined a number of each conceptual node connected with other nodes: 7 brown edges are the incident edges of concept pairs occurred in two domains, 13 green edges in an LCA domain, and 12 red edges are in an economic domain.

The TransR model then detects each point of miscommunication by averaging the vectors of semantic concepts detected in each text chunk with the codebook. As illustrated in Fig. 5, we implement a web application to demonstrate detecting concepts by using the TransR model. The web application has two parts: user profile and background at the left, and question and answer at the right. In the example, we can detect 9 concepts in the question, and 8 concepts in the replied. Evidently, we can make a point of miscommunication caused by ambiguous term “economic cost” related to both LCA and economics. Therefore, the average vectors of the question and answer texts significantly differ from each other, reflecting cross-domain miscommunication.

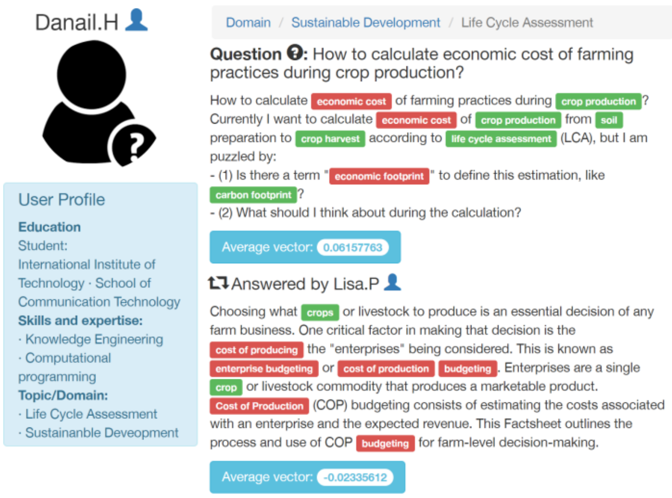

Fig. 5. An interface of a web application with a discussion forum for detecting concepts: nine detected concepts in the question text, eight detected concepts in the replied text, and the average vectors of the question and answer texts.

6 Conclusion and Future Work

In this paper, we present the cross-disciplinary approach integrating a method of Network Text Analysis and knowledge graph embedding to understand the relationship between questions and answers, in which knowledge is scattered as fragments.

Our approach can overcome research questions as follow. First, we analyze a discussion context by NTA’s workflow for extracting potential terminologies. Afterward, TransR is the vector-space model can estimate the positions of semantic concepts from two sources of knowledge. Under the extracted concepts, a concept may have multiple aspects relating different aspects of participants (i.e. domain experts and stakeholders). By so doing, we can extract the relationship, relevance, and consistency of each concept with respect to the discussion context.

With respect to the SD paradigm, our case study is a source of knowledge from the Q&A contexts under LCA) topic, represented in natural languages. All Q&A contexts existing height frequency of C^{CD} with the value of positional vectors are used to generate a co-occurrence networking visualization. The experimental result shows significant and consistent of communication comparing of questions and answers. Therefore, the main contribution of our research is to identify C^{CD} used across multiple domains that are the cause of miscommunication in domain-specific Q&A discussion.

In future work, we will integrate the cross-disciplinary approach to the collaborative framework [16] that is a communication space for enhancing a capability of discovering C^{CD} and detecting miscommunication.

Acknowledgment

This research is partially supported by Japan Advanced Institute of Science and Technology (JAIST), Japan, the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS) and by NRU grant at Sirindhorn International Institute of Technology (SIIT), Thammasat University and National Electronics and Computer Technology Center (NECTEC), Thailand. The authors are grateful to Wanwisa Thanungkano for sharing her LCA knowledge and experiences. LCA's materials and data are kindly provided by the LCA Laboratory, MTEC, Thailand.

References

1. Gruen, R.L., Elliott, J.H., Nolan, M.L., Lawton, P.D., Parkhill, A., McLaren, C.J., Lavis, J.N.: Sustainability science: an integrated approach for health-programme planning. Lancet. 372, 1579–1589 (2008).
2. Bernard, C.K.C., Anita, W.P.P.: Multidisciplinarity Interdisciplinarity and Transdis Ciplinarity in Health Research Services Education and Policy: 1. Definitions, Objectives, and Evidence of Effectiveness. Clin. Investig. Med. 29, 351–364 (2006).
3. Andresen, M.: Asynchronous discussion forums: Success factors, outcomes, assessments, and limitations. Educ. Technol. Soc. 12, 249–257 (2009).
4. Carley, K.M., Columbus, D., Azoulay, A.: AutoMap User's Guide 2012. Cent. Comput. Anal. Soc. Organ. Syst. (2012).
5. Daems, O., Erkens, M., Malzahn, N., Hoppe, H.U.: Using content analysis and domain ontologies to check learners' understanding of science concepts. J. Comput. Educ. 1, 113–131 (2014).
6. Aviv, R., Erlich, Z., Ravid, G., Geva, A.: Network analysis of knowledge construction in asynchronous learning networks. J. Asynchronous Learn. Networks. 7, 1–23 (2003).

7. Hecking, T., Hoppe, H.U.: A Network Based Approach for the Visualization and Analysis of Collaboratively Edited Texts. In: Proceedings of the First International Workshop on Visual Aspects of Learning Analytics co-located with 5th International Learning Analytics and Knowledge Conference. pp. 19–23 (2015).

8. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning Entity and Relation Embeddings for Knowledge Graph Completion. In: AAAI. pp. 2181–2187 (2015).

9. Alvargonzález, D.: Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences. Int. Stud. Philos. Sci. 25, 387–403 (2011).

10. Chaudhry, A.S., Higgins, S.: On the need for a multidisciplinary approach to education for knowledge management. Libr. Rev. 52, 65–69 (2003).

11. Bird Steven, Ewan Klein, and E.L.: Natural Language Processing with Python. O'Reilly Media, Inc. (2009).

12. Horrocks, I.: Ontologies and the semantic web. Commun. ACM. 51, 58–67 (2008).

13. European Union: Charter of fundamental rights of the european union. European Union, Brussels (2010).

14. International Organization for Standardization: ISO 14040 - Environmental management - Life Cycle Assessment - Principles and Framework. ISO, Geneva, Switzerland (2006).

15. Question Answering (Q&A) under topic; Life-Cycle Assessment (LCA) from ResearchGate website, A social networking site for scientists and researchers to share papers, https://www.researchgate.net/topic/Life-Cycle-Assessment.

16. Takhom, A., Ikeda, M., Suntisrivaraporn, B., Supnithi, T.: Toward Collaborative LCA Ontology Development: a Scenario-Based Recommender System for Environmental Data Qualification. In: Proceedings of EnviroInfo and ICT for Sustainability 2015. Atlantis Press, Copenhagen, Denmark (2015).

17. Horridge, M., Bechhofer, S.: The {OWL} {API}: {A} {Java} {API} for {OWL} ontologies. Semant. Web. 2, 11–21 (2011).

18. Glossary of economics: Glossary of Economics --- Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Glossary_of_economics.

19. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An Open Source Software for Exploring and Manipulating Networks, (2009).