A Study of Lexical Ambiguity in Large Forum Discussions for Multidisciplinary Knowledge Engineering

Akkharawoot Takhom[1,2], Prachya Boonkwan[3], H. Ulrich Hoppe[4], [1]Mitsuru Ikeda, Sasiporn Usanavasin[2], and Thepchai Supnithi[3]
[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi, Japan
[2]School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[3]Language and Semantic Technology Laboratory, National Electronics and Computer Technology Center, Pathum Thani, Thailand
[4]Department of Computer Science and Applied Cognitive Science within the Faculty of Engineering, University of Duisburg-Essen, Duisburg, Germany
Emails: [1]{akkharawoot, ikeda}@jaist.ac.jp, [3]{prachya.boonkwan, thepchai.supnithi}@nectec.or.th, [4]hoppe@inf.uni-due.de, [2]sasiporn.us@siit.tu.ac.th,

***Abstract*—Lexical ambiguity is a challenging issue in multidisciplinary knowledge engineering due to the tendency that lexical terms can be used among different domains with different specific meanings. Particularly in large forum discussions, such ambiguous cross-disciplinary terms hard to be identified and detected by the discussion participants because domain expertise from several relevant fields is required to detect those terms and discover the actual divergence of interpretation. Having many ambiguous terms in the discussion context will result in gradual misunderstanding and delayed knowledge construction. We studied the effects of data sizes and morphological analysis in discovering ambiguous cross-disciplinary terms in large forum discussions. Our findings are twofold. First, it is more likely to discover cross-disciplinary terms as forum discussions deepen. This correlates with domain experts' tendency to use general terms in metaphorically describing domain-specific concepts, therefore causing lexical ambiguity. Second, we found that lemmatization outperforms stemming in forming more understandable key terms. This is because lemmatization eliminates only inflectional affixes and keeps derivational affixes. On the other hand, stemming eliminates both types of affixes, causing semantic bleaching.**

Keywords—Multidisciplinary knowledge engineering; Lexical ambiguity, Network text analysis, Forum discussion, Morphological analysis

I. INTRODUCTION

In multidisciplinary knowledge engineering, domain experts have to communicate through the discussion context [1] in order to bridge the gap of fragmented knowledge [2]. As a hindrance of knowledge construction, there are three crucial issues that cause misunderstanding in a discussion exchange: lexical ambiguity, insufficient information in term coining, and out-of-scope topic shifts. In this paper, we focus on lexical ambiguity, where experts use general terms to convey different domain-related meanings.

One way to mitigate this issue is to detect an early sign of misunderstanding in the discussion context by identifying ambiguous cross-disciplinary terms with *Network Text Analysis* (*NTA*) [3]. NTA discovers an interrelationship sharing common terms in different disciplines using a text-mining method. The method extracts relationships between terms of possibly different categories from given texts and organizes these in the form of multi-partite networks. NTA has been employed in several applications, e.g., analyzing multidisciplinarity of knowledge management courses [4], exploring types of users in a discussion forum [5], and analyzing an understanding of science learners [6]. We believe that we can detect these ambiguous terms by associating them across domains as driven by a large amount of data.

In this paper, we will explore the use of NTA in detecting ambiguous cross-disciplinary terms in large forum discussions. We will examine the effects of using the NTA method to discover fragmented knowledge causing ambiguous cross-disciplinary terms in large forum discussions. Data preparation and preprocessing in the NTA workflow will be taken into account in the effects of data sizes and morphological analysis.

The rest of this paper is organized as follows. Section 2 defines fragmented knowledge in discussion contexts, and related works for analysis from a network perspective. Section 3 next introduces a methodology based on network text analysis: system overview and focus phases - data collection and data preprocessing. Section 4 then explains our case study on sustainable development. Afterward, Section 5 evaluates the experimental result and discusses significant issues. Finally, Section 6 concludes the paper.

II. BACKGROUND AND RELATED WORKS

A. Challenges in Multidisciplinary Knowledge Engineering

Multidisciplinary knowledge engineering [1] is characterized by the relevance of multiple academic disciplines in knowledge engineering. Due to its gnostic complexity, there is a communication bottleneck between domain experts, where

stakeholders have to collaborate with other participants in different perspectives and disciplines.

Particularly in learning fragmented knowledge [2], one particular challenge in conversational exchange is these experts often use general terms with different domain-related meanings. In large forum discussions, it is hard to identify ambiguous in cross-disciplinary terms and discussion participants cannot always detect those terms from several relevant fields on the fly. Many ambiguous terms in the discussion context will result in gradual misunderstanding and delayed knowledge construction. This issue is also known as *lexical ambiguity*.

B. *Network Text Analysis (NTA)*

Network Text Analysis (*NTA*) [3] is a text-mining method for detecting and encoding an interrelationship among terms from different categories constructing a network of the linked terms [7]. Fig. 1 shows a network of words analyzed by NTA.

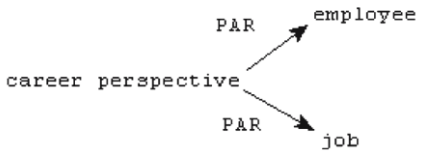


Fig. 1. An example of text analysis [7]: the concept "career perspective".

The ensuing semantic networks allow for identifying relationships between different categories. The categories are predefined in a "codebook" implementing a simple version of an ontology. This allows for capturing different knowledge categories as they appear in the multidisciplinary context.

NTA has been applied in various domains including social and educational studies. Chaudhry et al. [4] employed the NTA to analyze multidisciplinarity knowledge management courses that describe levels of courses, curriculum areas and topics, and differences in emphasis in teaching the courses in different departments and schools. Next, Hecking et al. [5] employed the NTA method to explore types of users in a discussion forum and analyze the visualize as a result using edited texts. Then, Daems et al. [6] employed the NTA method to analyze an understanding of science learners with discussion contents and domain ontologies.

In this paper, we propagate a data-driven approach to detecting lexical ambiguity in multidisciplinary forum discussion. Our approach draws on NTA as a basic method.

III. SYSTEM OVERVIEW

We closely follow the architecture of NTA described in [3], [6], [7]. As illustrated in Fig.2, a workflow of the NTA method consists of seven phases following the black arrows in the orange rectangles. Starting from the top part, contexts of questions and answer are observed in *Phase 1* (*Data Observation*). Then the selected knowledge sources are gathered in *Phase 2* (*Data Collection*) that we can use a data-handling tool for crawling data from a website and extract data structure.

In *Phase 3* (*Data Preprocessing*), the collected data are preprocessed including five techniques of *natural language processing*: (1) *tokenization* is to break text stream into meaningful terms, called tokens, (2) *removing stop words* for filtering out function words, such as 'the', 'is', and 'at.', (3) *morphological analysis* is to remove grammatical inflections from each term, and (4) *n-gram detection* for detecting a sequence of n adjacent elements from a string of tokens, such as bigram or 2-gram words. Pairs of words are counted by cumulative frequency.

There are yet two kinds of morphological analysis: lemmatization [8] (eliminating last inflectional affixes and preserving all derivational affixes) and stemming [9] (eliminating all affixes). For example, the word 'organizations' consists of the root *organ*, the derivational suffixes *-ize* (transforming a noun to a verb) and *-ation* (transforming a verb to a noun), and the inflectional suffix *-s* (adding plurality). Lemmatizing 'organizations' results in *organization*, while stemming the word will result in *organ*. Therefore, stemming is likely to cause semantic bleaching in most cases, because it may eliminate meaningful derivation affixes from the word. We will further study their effects in detecting ambiguous terms in Section IV.

As a next step, high-frequency terms are selected as potential terms in Phase 4 (potential-term selection) by determining the meaning of linguistic expressions in natural languages. In Phase 5, we exploit a semantic approach [10] to understand the semantic meaning. *Ontology* is a language for expressing conceptual knowledge and relations of the knowledge. In knowledge engineering, domain ontologies [11] capture the knowledge valid for a particular type of domain. The domain ontology is exploited to understand the semantic meaning of the fragmented knowledge [2].

In *Phase 6*, we prepare *a cross-domain codebook* to define relevant domains in the collected data by a triplet including selected potential terms, conceptual knowledge, and categories.

After preparing the collected data of a discussion context and a cross-domain codebook, *Phase 7* is to generate a network using a visualization tool for network generation.

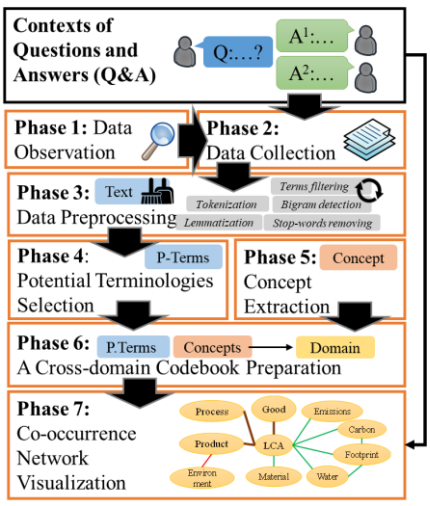


Fig. 2. An overview of a workflow of network text analysis [3], [12].

IV. CASE STUDY: SUSTAINABLE DEVELOPMENT

A. The Paradigm of Sustainable Development

The following case study shows the use of the approach in the multidisciplinary field of *Sustainable Development* (*SD*) [13], in which knowledge fragmentation [2] is an actual issue. SD addresses the sustainability of a natural system in general as well as techniques to support the sustainability of natural resources and ecosystem services. This paradigm related to many domains at least three main aspects: economy, social, and environment. To make valid contributions to this field, stakeholders need to consider aspects of all relevant domains.

In the case of *forum discussion*, stakeholders can share their knowledge with other relevant domain experts in replying to research questions in particular fields. This paper analyzes an environment aspect underlying the SD paradigm called *Life Cycle Assessment* (*LCA*) [14]. LCA is a guide to best practice in quantifying environmental resources used and released to the environment and evaluating opportunities to influence environmental improvements.

Nevertheless, ambiguous terms exist in communication contexts are taken into account in employing LCA knowledge, because the knowledge does not relate only a single discipline. Relevant stakeholders may not detect a lexical ambiguity or cannot address the fragmented knowledge explicitly. Sharing common terms by stakeholders are analyzed, and the NTA method is used to discover ambiguity in a large forum discussion.

B. Data Preparation

The experiment follows the NTA workflow (Fig.1). First, Data sources were observed (*Phase 1*) in several websites providing forum discussion that allows stakeholders to discuss in LCA topic. A social networking website, ResearchGate [15], provides forum discussion to share scientific publications, inquire and contribute information for scientists and researchers. As shown in Table I, the excerpt of communication contexts is analyzed manually: economic terms highlighted red italic, and LCA terms highlighted green italic.

TABLE I. AN EXCERPT OF COMMUNICATION CONTEXTS: 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)*.
Answer	Choosing what *crops* or livestock to produce is an essential decision of any *farm business*. One critical *factor* in making … the *cost of producing* the "enterprises" considered … 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 … *COP budgeting* for farm-level decision-making.

Contexts of forum discussion in this website were collected and extracted (*Phase 2*) by an open-source web-crawling platform, Scrapy [16]. The results [17] of Phase2 organized groups of data into two groups: 18,974 words on September 26, 2016, and 502,565 words on July 12, 2017. An excerpt of communication contexts: economic terms (red italic) and LCA terms (green italic).

C. Experiment Settings

After collecting data, this experiment designs data preprocessing (Phase 3) with three different NLP techniques denoted by techniques of natural language processing. As shown in Table II, the preprocessing results are compared in two data sizes using the WordNet lemmatization (WordNetLemmatizer) [8] and the Porter stemmer (PorterStemmer) [9].

TABLE II. A COMPARISON OF PREPROCESSING TECHNIQUES: LEMMATIZATION AND STEMMING IN TERM FREQUENCY (FRQ.)

Lemmatization				Stemming			
18,974 words	**Frq.**	**502,565 words**	**Frq.**	**18,974 words**	**Frq.**	**502,565 words**	**Frq.**
university	368	life	7,597	univers	372	life	7,597
lca	327	cycle	5,474	lca	339	cycl	5,546
data	249	system	3,338	product	263	product	4,165
life	242	life cycle	3,310	data	249	research	3,493
cycle	227	research	3,006	life	242	system	3,340
life cycle	219	time	2,866	cycl	227	life cycl	3,312
energy	191	assessment	2,834	life cycl	219	time	2,900
process	180	energy	2,353	energi	191	assess	2,848
technology	178	production	2,339	process	190	process	2,433
impact	177	process	2,284	technolog	178	energi	2,355

*Selected terms in Phase4

The experimental setting has six models: two data sizes (18,974 and 502,565) and cross-disciplinary codebook (data preprocessing and triplet sizes. Details of each model are set in Table III.

TABLE III. AN EXPERIMENTAL SETTING IN SIX MODELS.

Contexts Model		Cross-Disciplinary Codebook	
No.	**Word**	**Preprocessing**	**Triplet**
Model 1	18,974	R	291
Model 2	18,974	RL	288
Model 3	18,974	RS	165
Model 4	502,565	R	6,037
Model 5	502,565	RL	6,605
Model 6	502,565	RS	4,349
Note: Processing types R = Remove Stop Word, L=Lemmatization, S = Stemming			

In *concept extraction* (Phase 5), domain ontologies were chosen to examine LCA ontologies, named multidisciplinary LCA ontology (MLCA) [18]. Fig. 3 illustrates the excerpt of three upper concepts in the MLCA ontology: LCA concepts in a green circle, Life Cycle Costing (LCC) concepts in a blue circle and Data Quality Indicator (DQI) concepts in a yellow circle. 530 concepts were extracted and categorized into LCA domain. Whereas there is insufficient of a relevant domain, and we extend economic concepts by matching with a glossary from Wikipedia [19] containing 790 terminologies.

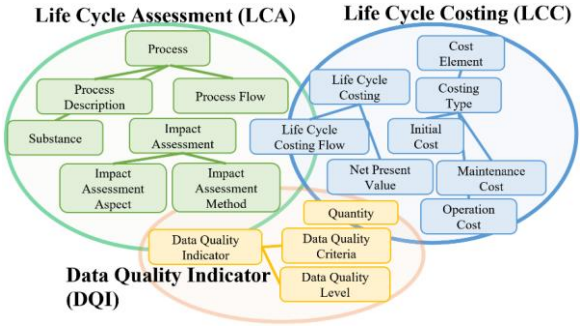


Fig. 3. An excerpt of three upper concepts in the MLCA ontology [18]: (1) LCA concepts in a green circle, (2) LCC concepts in a blue circle and (3) DQI concepts in a yellow circle.

The preprocessed data and domain concepts were used to construct a codebook (Phase 6), categorized depending on possible triplets. In the last phase, the collected data and a cross-disciplinary codebook are used to generate a co-occurrence network (Phase 7) by GePhi [20], a visualization tool for six different networks, as illustrated in Fig. 3.

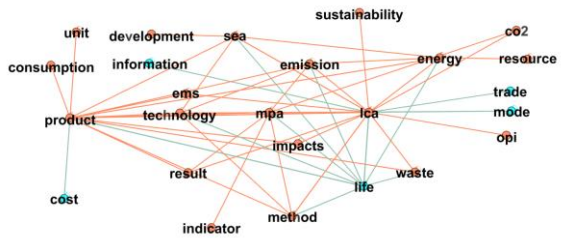
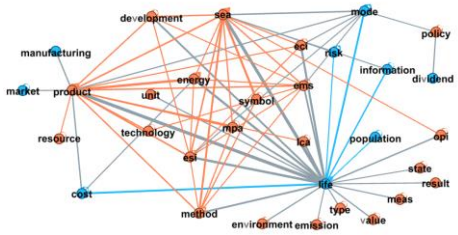
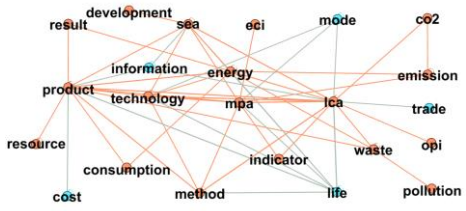
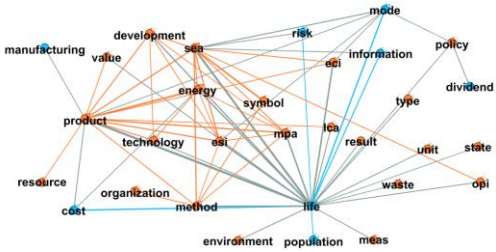
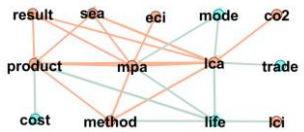
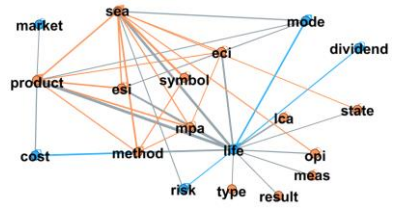
V. Result and Discussion

For the experimental result, research findings are discussed two significantly issues: effects of data sizes and effects of morphological analysis.

A. *Effects of Data Sizes*

It is more likely to discover the fragmented knowledge as forum discussions deepen. This correlates with domain experts' tendency to use general terms in metaphorically describing domain-specific concepts, therefore causing lexical ambiguity. At the last column of Table III, the six models are compared in three main criteria: (1) sizes of discussion contexts, (2) a codebook in pre-processing types and triplet sizes of a codebook, and (3) mapped results in nodes and edges. When the sizes of data are increased, we have a possibility to find ambiguous. ~~In parts of network visualization,~~ We take the different representation based on the designed model into account in an interpretation including usefulness and Significance of the relevant domains.

The experiment results in Figure 4 shows that we can discover ambiguous terms via NTA. Each edge is colored to represent co-occurrence in domains, i.e., orange for LCA, blue for economics, and gray for ambiguous cross-disciplinary

Data size: 18,974 words **Visualization setting:** Repulsion strange 2000; Attraction strange 0.1; Filtered edge weight at 10;	**Data size:** 502,565words; **Visualization setting:** Repulsion strange 2000; Attraction strange 0.1; Filtered edge weight at 100;
	
Model 1: Removing stop words (R)	**Model 4:** Removing stop words (R)
	
Model 2: Removing stop words and Lemmatization (RL)	**Model 5:** Removing stop words and Lemmatization (RL)
	
Model 3: Removing stop words and Stemming (RS)	**Model 6:** Removing stop words and Stemming (RS)

Abbreviations for processing types: R = Remove Stop Word, L=Lemmatization, S = Stemming

Fig. 4 Co-occurrence networks in six models designing in different data size, preprocessing techniques and cross-disciplinary codebooks.

terms. In Models 1, 2, and 3, the number of gray edges are quite low. When we incorporate more data, we observe the increase of gray edges in Models 4, 5, and 6, inferring that the more data, the more lexical ambiguity. Most of these nodes are general terms and usually convey metaphorical analogy across domains.

We observed the behaviors of misunderstanding as the discussion forum grows. In Figs. 1 and 4, we snapshot the same discussion forum in a duration of one year and plotted the term networks. In contrary to our preliminary belief, there is actually an increase of ambiguous terms in the communication exchange. We believe that this is due to topic shifts during the discussion, where common terms are used for metaphorical explanation.

B. Effects of Morphological Analysis

We found that lemmatization outperforms stemming in forming more understandable key terms. This is because lemmatization eliminates only inflectional affixes and keeps derivational affixes. On the other hand, stemming eliminates both types of affixes, causing semantic bleaching. As mentioned in a comparison [21], stemming can reduce all terms with the same stem to a common form whereas lemmatization can remove inflectional endings and returns the base form.

As shown in Table IV, the mapped nodes in each model are considered in two criteria: redundant nodes and semantically bleached nodes. Models 2 and 5 outperform the others in both criteria. In the smaller data size, Model 2 completely eliminates both redundant nodes and semantic bleaching. So does Model 5 in the larger data size.

The results suggest that only removing stop words is inadequate for detecting ambiguous cross-disciplinary terms, because redundant nodes are generated in the text network. It also implies that lemmatization generates more meaningful terms than stemming does, because derivational suffixes are crucial in forming the domain concepts. That explains why stemming, eliminating all suffixes, results in semantically bleached nodes.

TABLE IV. A COMPARISON OF THE EXPERIMENTAL RESULTS.

Contexts Model		**Mapped Result**			
No.	**Word**	**Node**	**Redundant Node**	**Semantically bleached nodes**	**Edge**
Model 1	18,974	81	2	0	1,273
Model 2	18,974	86	**0***	**0***	1,302
Model 3	18,974	45	0	8	438
Model 4	502,565	272	23	0	10,865
Model 5	502,565	250	**0***	**0***	9,566
Model 6	502,565	114	0	23	2,480

VI. CONCLUSION AND FUTURE WORK

In this paper, we examined the effects of data sizes and morphological analysis in discovering ambiguous cross-disciplinary terms in large forum discussions using Network Text Analysis (NTA). We used sustainable development as our case study.

Our findings are two folds. First, domain experts tend to explain domain-specific concepts with general words, causing ambiguous cross-disciplinary terms. The longer the forum discussion becomes, the more ambiguous terms are introduced. Second, lemmatization is a better-preprocessing step than stemming, because it corresponds to the experts' formation of domain concepts.

Our future work remains as follow. First, we plan to choose domain concepts with better criteria than term frequency, because it neglects crucial but infrequently mentioned key terms. Second, we plan to automate the construction of codebooks from large forum discussions to reduce human labor. Third and finally, we plan to take into account some of consecutive words as domain concepts and investigate their effects on NTA.

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] D. Alvargonzález, "Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences.," *International Studies in the Philosophy of Science*, vol. 25, no. 4, pp. 387–403, Dec. 2011.

[2] E. Becker, T. Jahn, I. Stieß, and P. Wehling, *Sustainability: A cross-disciplinary concept for social transformations*. Unesco, 1997.

[3] J. Diesner and K. M. Carley, "Using Network Text Analysis to Detect the Organizational Structure of Covert Networks * Jana Diesner, Kathleen M. Carley," in *Communication*, 2004.

[4] A. S. Chaudhry and S. Higgins, "On the need for a multidisciplinary approach to education for knowledge management," *Library Review*, vol. 52, no. 2, pp. 65–69, Mar. 2003.

[5] T. Hecking and H. U. Hoppe, "A network based approach for the visualization and analysis of collaboratively edited texts," in *CEUR Workshop Proceedings*, 2015, vol. 1518, pp. 19–23.

[6] O. Daems, M. Erkens, N. Malzahn, and H. U. Hoppe, "Using content analysis and domain ontologies to check learners' understanding of science concepts," *Journal of Computers in Education*, vol. 1, no. 2–3, pp. 113–131, 2014.

[7] R. Popping, "Knowledge graphs and network text analysis," *Social Science Information*, vol. 42, no. 1, pp. 91–106, 2003.

[8] H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor, "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text," *Journal of biomedical semantics*, vol. 3, no. 1, p. 3, 2012.

[9] P. Willett, "The Porter stemming algorithm: then and now," *Program*, vol. 40, no. 3, pp. 219–223, 2006.

[10] I. Horrocks, "Ontologies and the semantic web," *Communications of the ACM*, vol. 51, pp. 58–67, 2008.

[11] R. Studer, V. R. Benjamins, D. Fensel, and others, "Knowledge engineering: principles and methods," *Data and knowledge engineering*, vol. 25, no. 1, pp. 161–198, 1998.

[12] A. Takhom, P. Boonkwan, M. Ikeda, S. Usanavasin, and T. Supnithi,

"Reducing miscommunication in cross-disciplinary concept discovery using network text analysis and semantic embedding," in *CEUR Workshop Proceedings*, 2017, vol. 2000, pp. 20–31.

[13] European Union, *Charter of fundamental rights of the european union*, vol. 56. Brussels: European Union, 2010.

[14] A. Astrup Jensen *et al.*, *Life Cycle Assessment - A guide to approaches, experiences and information sources*. EEA (European Environment Agency), 1997.

[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," 2017. [Online]. Available: https://www.researchgate.net/topic/Life-Cycle-Assessment. [Accessed: 10-Sep-2016].

[16] D. Myers and J. W. McGuffee, "Choosing Scrapy," *Journal of Computing Sciences in Colleges*, vol. 31, no. 1, pp. 83–89, 2015.

[17] "Question Answering (Q&A) under topic; Life-Cycle Assessment (LCA) from ResearchGate website, A social networking site for scientists and researchers to share papers," 2016. [Online]. Available: https://www.researchgate.net/topic/Life-Cycle-Assessment. [Accessed: 10-Sep-2016].

[18] A. Takhom, S. Usanavasin, T. Supnithi, and M. Ikeda, "Collaborative ontology development approach for multidisciplinary knowledge: A scenario-based knowledge construction system in life cycle assessment," *IEICE Transactions on Information and Systems*, vol. E101D, no. 4, pp. 892–900, 2018.

[19] Glossary of economics, "Glossary of Economics --- Wikipedia, The Free Encyclopedia," 2016. [Online]. Available: https://en.wikipedia.org/wiki/Glossary_of_economics. [Accessed: 10-Sep-2016].

[20] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks." 2009.

[21] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: a comparison of retrieval performances," 2014.