International Journal of Smart Computing and Artificial Intelligence
International Institute of Applied Informatics
2019, Vol. 3, No. 1, 71 - 86

Construction of Urban Problem LOD using Crowdsourcing

Shusaku Egami [*], Takahiro Kawamura [†], Kouji Kozaki [‡], Akihiko Ohsuga [*]

Abstract

Municipalities in Japan have various urban problems such as traffic accidents, illegally parked bicycles, and noise pollution. However, using these data to solve urban problems is difficult, as these data are not structurally constructed. Hence, we aim to construct the Linked Data set that will facilitate the solving of urban problems. In this paper, we propose a method for semi-automatic construction of Linked Data with the causality of urban problems, based on Web pages and open government data. Specifically, we extracted causal relations using natural language processing and crowdsourcing to include problem causality in the Linked Data. Then, we provided an example query to confirm the relationships be-tween several problems. Finally, we discussed our crowdsourcing task design for extracting urban problem causality.

Keywords: Linked Open Data, Crowdsourcing, Urban problem, Causal Relation Extraction.

1 Introduction

Japanese local governments have multiple urban problems such as traffic accidents, illegally parked bicycles, noise pollution, graffiti etc. Local governments have been discussing countermeasures to solve these urban problems. Since a data-driven approach to solve these urban problems has attracted attention, open data published by municipalities are increasing.

The World Wide Web Consortium (W3C) recommends that the Open Data should be structured according to the Resource Description Framework (RDF) and the relation links should be created between the data elements. These data are called Linked Open Data (LOD). We believe that structured data related to classification and causality of urban prob-lems, countermeasures of municipalities, and citizens' voice are required to solve urban problems. Therefore, we aim to build an analysis infrastructure of urban problems by designing schemata and storing data as LOD. This infrastructure can help predict the extent of impact of the urban problems by tracing the causality and their hierarchical

* The University of Electro-Communications, Tokyo, Japan
† Japan Science and Technology Agency, Tokyo, Japan
‡ Osaka University, Osaka, Japan

links, and also help to consider the countermeasures that are proposed by the municipalities. In this paper, we designed a data schema representing urban problems causality. Then, we propose a method for semi-automatically extracting causalities of urban problems using natural language processing (NLP) and crowdsourcing. Finally, as a use case of the resulting LOD, we provide an example query to confirm the relationships between several problems. Our contributions are as follow:

1. Designing a schema of urban problem causality;
2. Proposing a method for semi-automatically constructing Linked Data with causality;
3. Publishing the data on the web as LOD; and
4. Presenting an example query to consider solutions of urban problems

The remaining sections of this paper are organized as follows. In Section 2, an overview of LOD related to urban problems and city data is described. In Section 3, the schema design is described. In Section 4, the method for constructing LOD related to urban problems with causality is described. In Section 5, an example query is presented, and we discuss the solutions of urban problems. In Section 6, we discussed the extraction results, crowdsourcing results, and the designed schema. Finally, Section 7 concludes this paper with some feasible future extensions.

2 Related Work

2.1 Knowledge Graph for solving social issues

Some studies have proposed use of linked data for solving social issues. Szekely et al. [1] built linked data from crawled sites for combating human trafficking and developed the lost children search system. The system has been deployed by six law enforcement agencies and several NGOs. Szekely et al. built linked data related to a specific domain of social problems, whereas this paper aims to build LOD related to multiple urban problems with causality.

In our previous work [3], we built and visualized LOD for solving illegally parked bicycles, which is one of the urban problems that needs to be resolved in Japan. In addition, we proposed a methodology for designing LOD schema of the urban problems that are occurring on a daily basis, such as illegally parked bicycles. In this methodology, all the steps were accomplished manually. Since the number of web documents collected was limited and the task relied on the knowledge of workers, the previous approach suffered from low coverage with respect to the extraction of the causality of urban problems. Moreover, since the constructed LOD was based on the schema extended from Event Ontology[1], and thus it was difficult to search urban problem causality using OWL inference rules. Thus, in this paper we defines a new LOD schema representing urban problem causality.

Shiramatsu et al. [4] proposed LOD to share goals to solve social issues. A goal-matching method using LOD has been proposed for facilitating civic technology (Civic Tech). The Civic Tech is aimed at solving social issues using information technology through collaboration between citizens and local governments.

[1] http://motools.sourceforge.net/event/event.html

Furthermore, it has been reported that a web application called GoalShare was developed and applied in domestic Civic Tech events. However, Shiramatsu's LOD mainly describes the public goals for solving social issues and does not describe causalities. Associating them with the LOD proposed in our study will facilitate social problem solving in the future.

2.2 Knowledge Graph for analyzing city indicators

Santos et al. [5] defined city knowledge graph as owl in order to analyze various city indicators. They proposed the quality of experience (QoE) Indicators Ontology representing calculated numerical values and supporting convenient visualization. The dashboard application, which can generate widgets visualizing knowledge graph data was also developed. Pileggi et al. [6] defined the ontological framework and implemented it as OWL-DL ontology to represent dynamic fine-grained urban indicators. In order to simplify the understanding of the data structure as well as the facilitation of its usability, the ontology was partitioned in five sub-ontologies; Indicator, Data, Profiling, Computations and Geographic Context based on the function of the scope within the model.

LinkedSpending [7] are linked data based on OpenSpending[2], which is an open platform for public financial information, including budgets, spending, balance sheets, procurement, etc. As of October 2017, 2,266 data sets from 77 countries have been registered. The data is modeled based on RDF Data Cube vocabulary[3], which is designed for modeling multidimensional data such as statistical data. However, these linked data do not describe urban problems and cannot be directly used to solve urban problems.

2.3 Crowdsourcing and NLP for Linked Data

Demartini et al. [8] proposed an entity linking method using crowdsourcing. Crowdsourcing was used to improve the quality of the links, and they developed the probabilistic framework to integrate inconsistent results. Celino et al. [9] developed a mobile application to link Point of interests (POIs) data to pictures using crowdsourcing. They introduced the method of game with a purpose (GWAP) [10] to give users incentives. However, there is no study to construct LOD related to urban problem causality using crowdsourcing.

Nguyen et al. [13] proposed a method for constructing Linked Data concerning users' activities. Conditional Random Field (CRF) was used to extract the users' activities from Japanese weblogs, and then triples related to *action*, *object*, *time* and *location* were constructed. This Linked Data would be applied to analize users' activities at the time of an earthquake. LODifier [14] also extracted entities from unstructured text using a Named Entity Recognition (NER) system Wikifier [15], and combined the entities to DBpedia and WordNet. There are many other studies using NLP techniques to construct Linked Data sets. However, the accuracy of the methods to extract the necessary words from unstructured documents in those studies is still unsatisfiable. Thus, in this paper, we combined an NLP technique and crowdsourcing in order to extract urban problem causality.

There are many study for extracting causality. Sakaji et al. [16] proposed a method for extracting causality from unstructured text using clue phrases such as "*tame*: because" and "*niyori*: due to". Then, Ishii et al. [17] constructed causal network using clue phrases to support understanding of news. We first tried to extract causality using clue phrases.

[2] https://openspending.org
[3] https://www.w3.org/TR/vocab-data-cube/

However, since many causal relationships other than urban problems were extracted, we used using Japanese dependency analysis [11] and crowdsourcing.

3 Designing a schema of problem causality

In the previous work [2, 3], we collected data related to illegally parked bicycles, which is one of the urban problems in Japan, and have built the LOD for solving that problem. In addition, we have proposed a methodology for designing LOD schema of the urban problems that is occurring on a daily basis, such as illegally parked bicycles as follows:

1. Extraction of domain requirements
    a. Select an ontology that models the urban problem
    b. Search for articles on the urban problem using a search engine
    c. Extract keywords from the articles based on properties of the ontology
    d. Cluster the keywords
2. Designing schema
    a. Design classes based on the ontology
    b. Design instances and properties based on the clustering results

An existing ontology was selected in order to build LOD based on ontology. Next, the worker searched about the urban problem using search engines like Google. Then, the worker investigated the top 10 articles and their references, and manually extracted the keywords based on the properties of the existing ontology. Keywords that are not defined in the ontology, but appear to be important in the article were also extracted. The extracted keywords were clustered manually. Finally, their instances were designed in reference to the grouping result. In this methodology, all the steps were accomplished manually. Also, since the number of Web documents collected was small and the task depends on the knowledge of one worker, this approach suffered from low coverage with respect to the extraction of the causality of urban problems. In addition, all existing ontology were not suitable to causal inference.

Therefore, in this paper, we first propose a schema representing urban problem causality, then semi-automatically extract urban problems causality from web documents. Our LOD is mainly for investigation of solutions to urban problems. Specifically, querying the LOD enables local governments to consider the solutions. Thus, we designed the LOD schema shown in Fig. 1 to represent urban problem causality. All resources are classified as UrbanProblem or NotUrbanProblem. There are two main causality properties, upv:factor and upv:affect. Both properties are owl:TransitiveProperty and are sub-properties of the upv:related property. Since all urban problems are not events that have time and spatial thing, we did not reuse the event:factor property in the Event Ontology. Also, there are sub-properties of upv:factor and upv:affect to represent agreements of crowdsourcing. By dividing the causality properties into upv:factor and upv:affect, it is possible to reason forward or backward chainings with agreement levels restricting domain or range. For example, when users extract the strong causality, the upv:factor_level4 and the upv:affect_level4 properties can be used, and when users extract the causality regardless of the agreement, the upv:factor and the upv:affect properties can be used in SPARQL queries.

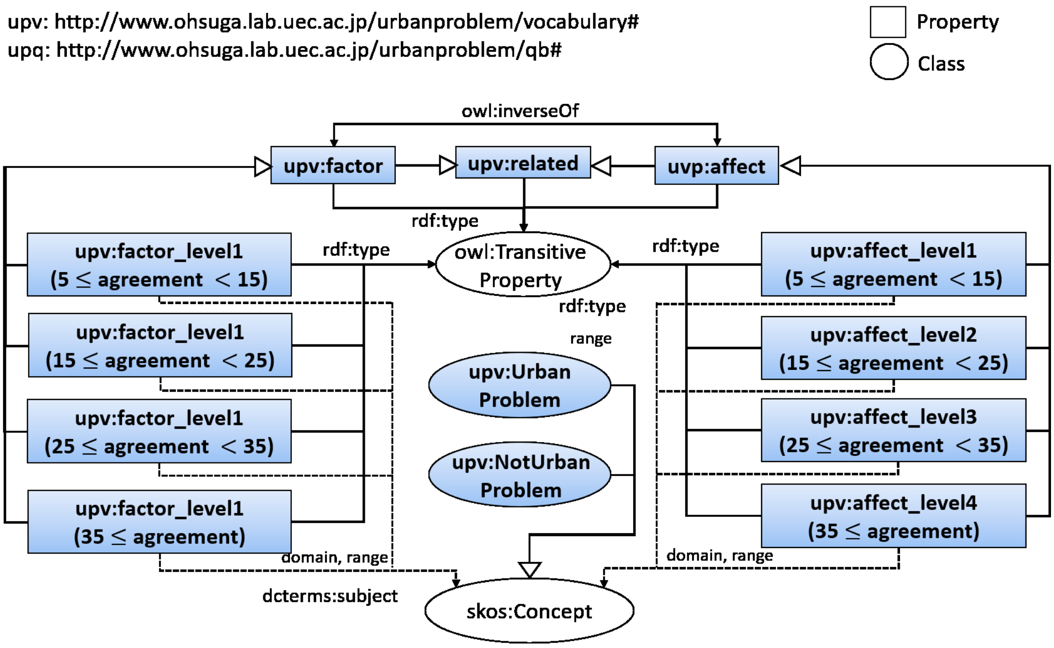

Figure 1: Schema of urban problem causality

4 Semi-automatic construction of LOD using NLP and Crowd-sourcing

4.1 Extraction of causality words using NLP

In this section we propose a method of semi-automatically extracting the causality of urban problems as follows.

(1) Collect web documents using a search engine.

(2) Extract causality words from the collected documents using NLP.

(3) Generate word clouds based on the extracted words.

(4) Filter the extracted words using crowdsourcing.

We collected documents by search engines using names of urban problems and synonyms of “factor” as keywords. For example, the first keyword is “noise,” and the second keyword is “factor” and its synonyms, such as “element”, “origin”, and “cause”. We obtained the synonyms of the second keyword from Japanese WordNet[4]. Then, we obtained the document lists using Google Custom Search API[5] and Bing Web Search API[6]. We separately collected 50 HTML files and 50 PDF files for each keyword sets (different com-binations of the first and second keywords). We collected HTMLs and PDFs separately to

[4]http://compling.hss.ntu.edu.sg/wnja/index.en.html
[5]https://developers.google.com/custom-search/?hl=en
[6]https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

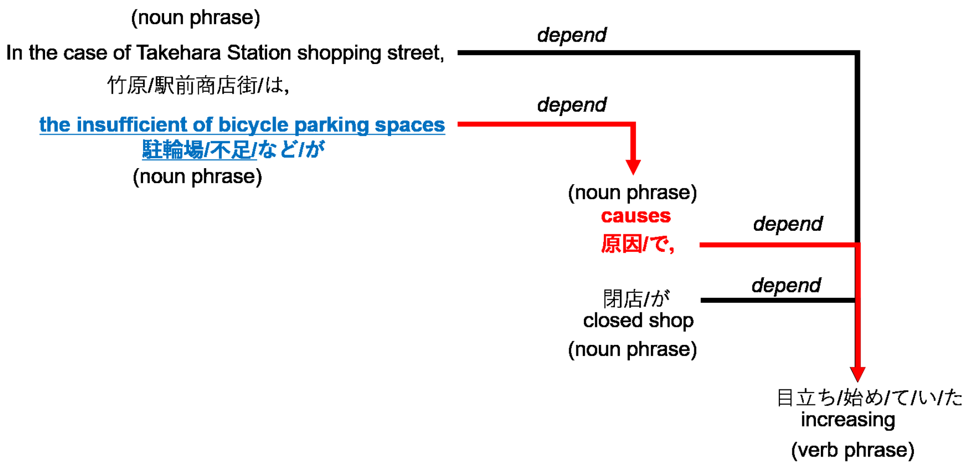

Figure 2: Extraction of candidates for causality words

collect reports published by governments and citizens' voices. However, many unrelated documents were collected in this step; thus, the documents containing few words related to urban problem name were excluded.

Next, we extracted noun words using Japanese morphological analysis. In order to facilitate the subsequent crowdsourcing process, we concatenated the verbal nouns and constructed noun phrases. For example, the phrase "preventing delinquency" was split into "preventing" and "delinquency" using morphological analysis, but these words were concatenated to a noun phrase in this study.

Then, we extracted noun phrases that have causal relationships to the synonyms of the "factor" based on dependency relations in each sentence using Japanese dependency analysis [11]. Figure 2 shows the procedure of this method. In this case, since "駐輪場不足 (the insufficient of bicycle parking spaces)" is noun phrase and directly depend the phrase including the synonym of "factor", we extract the "駐輪場不足" as a candidate for causality word.

Likewise, words related to the influence of urban problems were extracted, using the synonyms of "influence" as the second keywords, such as "affect", "effect", and "evoke".

4.2 Filtering causality words using crowdsourcing

We generated word clouds based on extracted possible causality words and filtered the words using crowdsourcing. Since assumed that the word clouds raise the impression of words and make it easier to extract important words, for workers than a simple list.

First, similar words obtained in Section 4.1 were integrated by edit distances. We used Jaro-Winkler distance [12] to calculate the similarity of words, and the threshold was empirically set to 0.8. When similar words were found, the number of occurrences of words was integrated to the longest word. Figure 3 and Figure 4 show word clouds of "noise". When the frequency of a specific word was high, the word size became large and was placed close to the center of the cloud. The color of words was set randomly. Then, we ordered two tasks for crowdsourcing: "Select 10 words that are considered factors of noise pollution" and "Select 10 words that are considered influence affected by noise pollution."

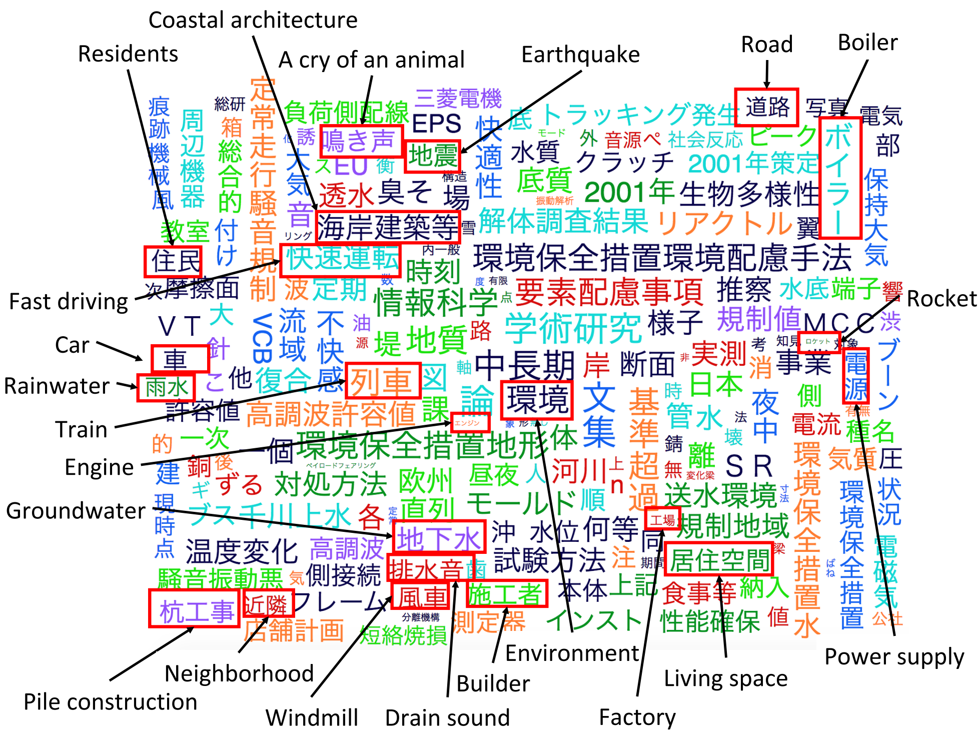

Figure 3: A word cloud of noise factors

In this paper, we used Lancers[7] as the crowdsourcing service. We set the reward for the two tasks at 50 JPY, and asked up to 50 people to work on each problem. Then, we took words that were selected by more than 10% of the workers. The selected words were those translated in Fig. 3 and Fig. 4.

4.3 Building LOD based on the extracted causality words

We built LOD based on designed schema using the extracted words. Since the Lancers can export the results of crowdsourcing in a CSV format, we converted the CSV to an RDF file based on the designed schema using Apache Jena[8]. Specifically , urban problem resources were created as sub-classes of the upv:UrbanProblem class, and others were created as sub-classes of the upv:NotUrbanProblem class. The causality links were created using the upv:factor and the upv:affect properties corresponding to the number of the agreement. In addition, we extracted noun words from the name of each class and created hyper classes based on them.

Figure 3 shows part of the LOD finally constructed in this study. The resulting LOD is accessible in our website[9]. There are 49,386 triples in the LOD. We validated our LOD using RDFUnit [18], which is a test driven data-debugging framework. The 68 test cases were automatically generated, and then all test cases passed. There were no timeout, no error, and no violation instances. Therefore, the result showed that we

[7] http://www.lancers.jp

[8] https://jena.apache.org/

[9] http://www.ohsuga.lab.uec.ac.jp/urbanproblem/

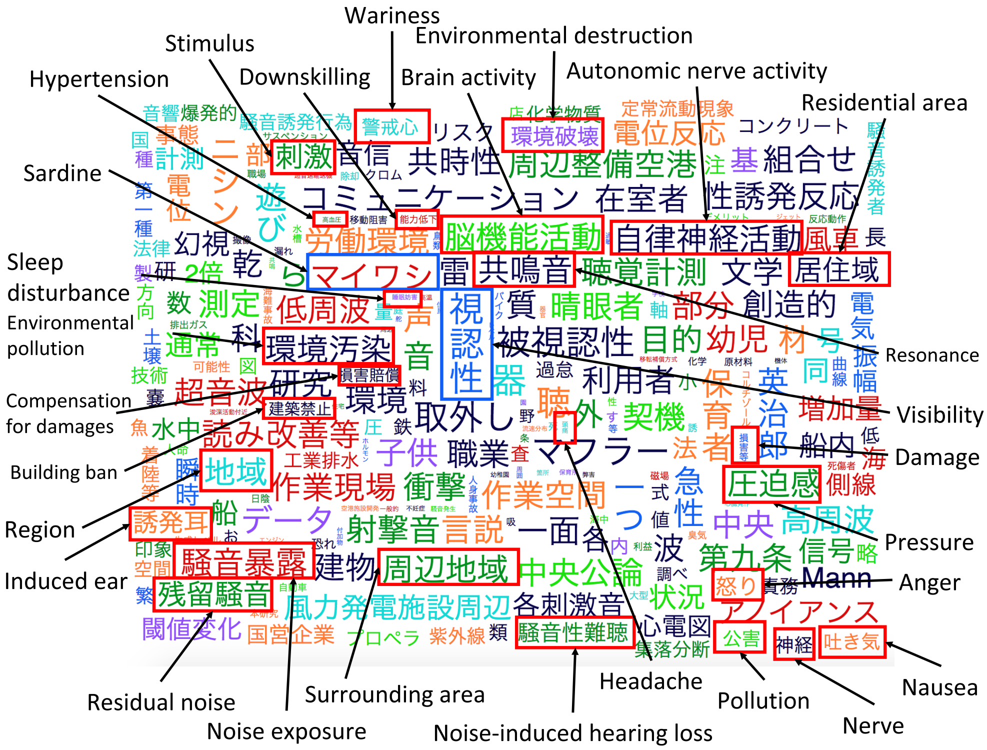

Figure 4: A word cloud of noise influences

correctly reused existing vocabularies without violation of domains and ranges restrictions. All resources are linked, and there is no independent resources.

4.4 Example Query

Figure 7 shows part of graph extracted from our LOD using a SPARQL query of searching causality of Littering as follows.

```
PREFIX upr: <http://www.ohsuga.lab.uec.ac.jp/urbanproblem/resource#>

PREFIX upv: <http://www.ohsuga.lab.uec.ac.jp/urbanproblem/vocabulary#>

SELECT DISTINCT ?influence ?factor

FROM <http://www.ohsuga.lab.uec.ac.jp/urbanproblem>

WHERE {

    upr:Littering upv:affect+ ?influence ;
    upv:factor+ ?factor .

}
```

In this figure, we found that insufficient `Installation of ashtrays` affects the `Littering`, and the `Littering` affects the `Groundwater pollution`. Thus, the direct

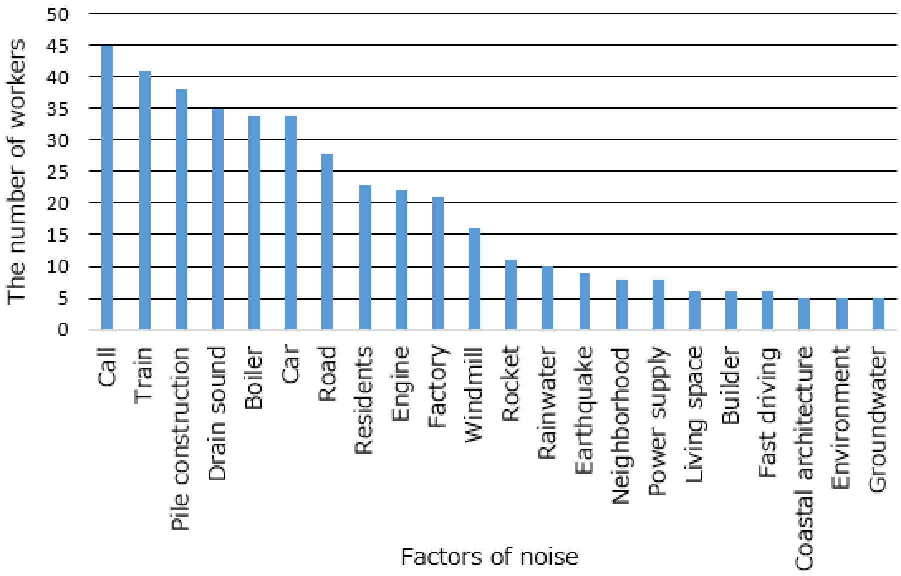

Figure 5: Selected words from Fig. 1 using crowdsourcing

relation with the `upv:affect` property from the `Littering` to the `Groundwater` is inferred by our predefined rule that an influence caused by a thing propagates to a super concept. Furthermore, the causality extraction for the `Noise` indicated that the `Groundwater` also affects the `Noise`. As the result, we found that as a chain reaction the insufficient `Installation of ashtrays` may affect `Littering`, the `Groundwater pollution`, and also the `Noise` according to the transitive relations of the `upv:affect` property. In fact, there is a case that cigarettes littered by street smokers flow into the groundwater; thus, the groundwater is polluted, and then the water flow becomes noisy, and finally, the plumbing to improve the water quality causes the further noise. Therefore, we can suggest that strengthening countermeasures for solving street smoking may lead to effective solutions for these serial problems since `Littering` is located at upstream side of the problems.

5 Discussion

5.1 Discussion of causality words extraction

Table 1 shows the statistics of causality words extraction. There were documents, which have been deleted and are machine-unreadable, and we could not use these documents. Also, documents unrelated to urban problems were included when a second keyword adversely effect on search. Specifically, medical documents containing a description of causality of disease were mistakenly collected. In the future, we consider using a key phrase with quotation marks such as "騒音の原因 (factor of noise)".

The omission of words related to urban problems was a cause of lowering the agreement in a process of causality word extraction process using morphological analysis and Japanese dependency structure analysis. We found that there are cases, in which the chunk extracted by our method does not match another chunk describing causality. Since there are many complex sentences in the documents published by governments,

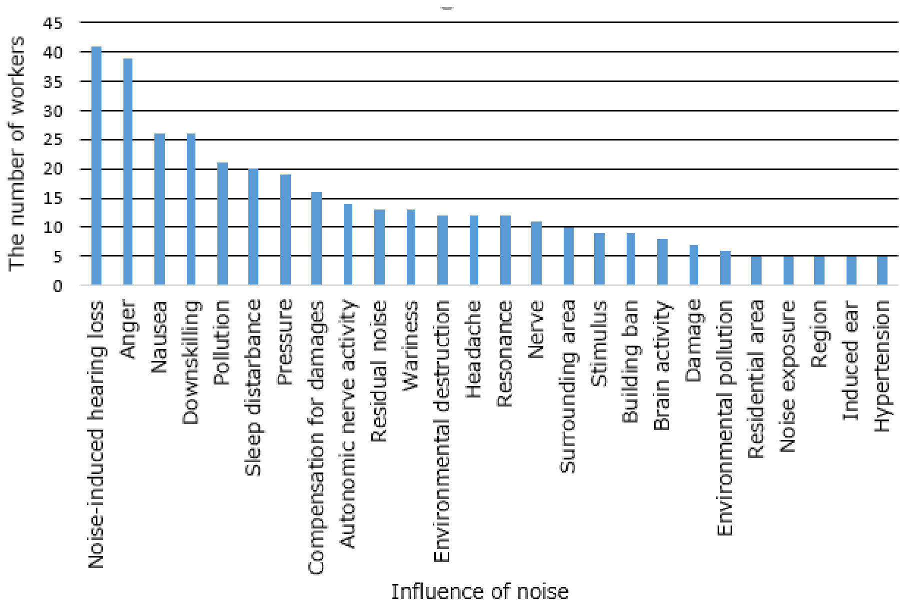

Figure 6: Selected words from Fig. 2 using crowdsourcing

Table 1: Statistics of causality words extraction

	# of documents including urban problem words	# of sentences including synonyms of "factor" and "influence"	# of extracted words
Factor	1,438	4,481	3,110
Influence	2,465	9,082	4,661

we could not extract the chunks containing causality words in many cases. To solve this problem, we will define extraction rules in future, such as extracting chunks depending other chunks containing the urban problem names.

In addition, there were words that had unknown relevance to each urban problem, and the words gathered in the center of the word cloud obstructed the selection of correct causality words. For example, in Fig. 4, the words which had unknown relevance to noise were clustered in the center of the word cloud, such as "視認性 (Visibility)" and "マイワシ(Sardine)". The factor, "Visibility", was described in documents related to remodeled cars, and it was mentioned that a colored film to a car window affected visibility. In the same documents, it was mentioned that removal of muffler leads to increase the noise. The factor "Sardine" was described in the documents related to auditory responses of fish, and it was mentioned that the low-frequency noise affects fish behavior. Therefore, there were cases, where the causality words were extracted from another part of the sentence although there were sentences containing "noise". Also, there were cases where causality words were extracted from descriptions not related to the urban problem. To exclude these errors, we will extract words that commonly appear in multiple documents, instead of words that appear many times in a single document.

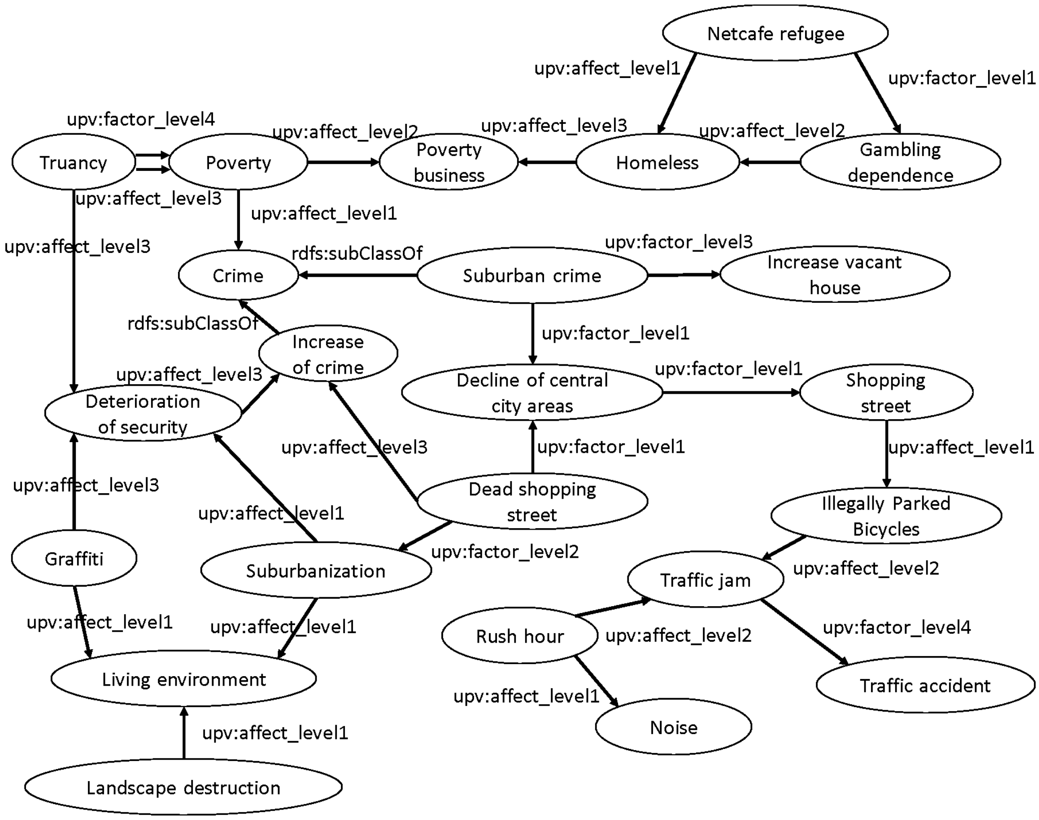

Figure 7: Part of constructed LOD

5.2 Discussion of crowdsourcing

To calculate the agreement of causality word selection in the crowdsourcing, we used Fleiss' s kappa [19]. The number of users was 50. The average number of extracted words related to urban problem factors was 0.291, and the number of words related to influence was 0.212. The total average of agreement was 0.256, and it was fair agreement according to the benchmark [20]. Therefore, the extraction method of the causality word and the display of the word cloud can be considered to be appropriate to some extent.

Especially , the agreement of factor of traffic accidents was quite high, since the various instances of traffic accidents being reported by the Metropolitan Police Department, educational institutions, and news organizations resulted in the workers having extensive background knowledge. On the other hand, the high agreement of noise factors is because the workers have had the experience that affected by noise in their daily life; thus, the word displayed in the word cloud was easy for them to associate.

In this study, we conducted 17 crowdsourcing experiments. We set the maximum number of workers to 50, and we received complete responses within 24 hours in all tasks. The number of unique workers was 287, and the number of repeaters was 141. Figure 9 shows the changes of the number of repeaters. The number of the repeaters has increased rapidly, and was over 35 since the sixth experiment. The maximum number of repeaters was 46. In the Lancers, workers can follow clients like SNS. The number of followers of our account increased to 58 through this study. Thus, 58 followers always received notifications when

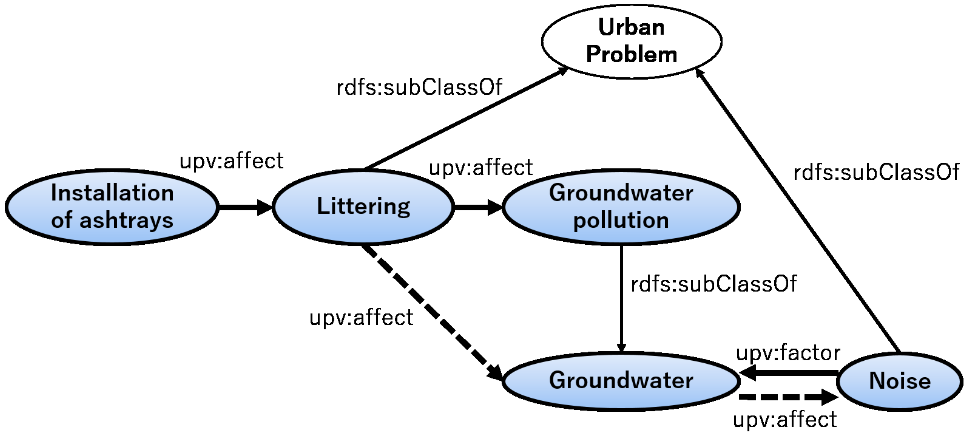

Figure 8: Causality of littering

we created tasks. From these results, we consider that our crowdsourcing task design was appropriate, since this number of followers are high in the tasks in the Lancers.

We provided a form for remarks at the end of each task. Table 2 shows the workers' comments. As a result, the number of opinions including causality and solutions related to urban problems was the largest, except for "Nothing". These comments included the reasons why the workers selected that causality words, other candidates for causality words, actual experiences related to the urban problems, and solutions. For example, we obtained a comments as follows:

- Traffic jam is sometimes caused by not only road conditions but "road rage" which is retaliatory behavior exhibited by a driver including dangerous driving and verbal insults.
- The largest factor of graffiti is rebellion to society, and the largest influence of graffiti is deterioration of social security.
- The littering is very danger because car accidents are caused by stepping on garbage on a rainy day. Also, I think that people litter at ease when there is garbage on the roads.

These comments have some causality words not included in the word clouds. Thus, we are considering to add them to our LOD these comments in future. We classified them as high-level awareness workers related to urban problems.

The ratio of impressions related to urban problems was 1.8%. These comments included dissatisfaction with urban problems. We classified them as middle-level awareness workers related to urban problems. By adding questions to our crowdsourcing tasks, there is a possibility that we can obtain additional opinions related to the urban problems from these workers. Also, there is a possibility that middle-level awareness workers become high-level awareness workers.

The ratio of impressions of the tasks were 3.8%. There were some comments that the task was difficult or the task was interesting. We classified them and respondents of "Nothing" as low-level awareness workers related to urban problems. We consider

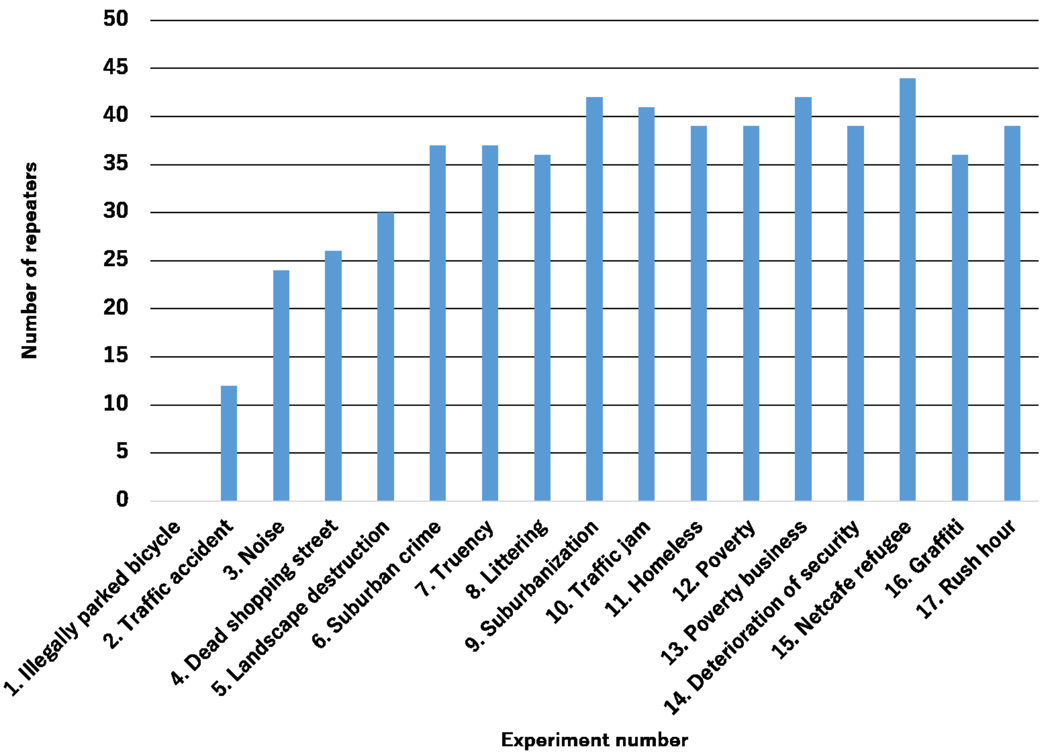

Figure 9: The number of repeaters of crowdsourcing

that their works affect the lowering the agreement of crowdsourcing. Therefore, we should consider methods to train the low-level awareness workers in the future as in [21].

5.3 Discussion of schema design

Regarding the instance parts of the designed LOD schema, there are some instances for which we could not obtain real data, for example, time, place, and name of poverty businesses and dead shopping streets. Therefore, although there is a possibility that it can be used for causality analysis at the class level, but it is currently difficult to use it for the purpose of causality analysis at the instance level. In the future, negotiations with municipalities are necessarily for collecting and publishing the data.

6 Conclusion

In this paper, in order to analyze urban problems, we extracted the causality words of urban problems from web documents and open government data, filtered the results using crowdsourcing, and then constructed LOD including causalities of urban problems. Through the example query, we confirmed that the constructed LOD can contribute to considering countermeasures to the problems. In this study, the urban problem causality was extracted based on government reports, research results of the sociology domain, human knowledge, facts, and association of ideas. Thus, we can explain the causality by confirming the source data.

Table 2: Classification of workers' descriptions in the remarks column

Comment	Amount	Ratio
Opinions including causality and solutions related to urban problems	82	9.6%
Impressions related to urban problems	16	1.8%
Impressions of the task including difficulty	33	3.8%
Nothing	719	84.5%

We will increase intermediate nodes in the case that it is difficult to understand the causality by improving the extraction method.

In the future, we will focus on other urban problems. Also, we will collect city data, such as countermeasures to urbanproblems and citizen's voice from open data and SNS, and add those data to our LOD. In addition, since one of the authors of this paper is a member of this council, we are now cooperating with the Osaka City Citizen Bureau through Osaka City Citizen Activity Promotion Council to consider several urban problem solutions based on our LOD.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP16K12411, JP16K00419, JP16K12533.

References

[1] P. Szekely, C. A. Knoblock, J. Slepicka, et a, "Building and Using a Knowledge Graph to Combat Human Trafficking," In: Proceedings of the 14th International Semantic Web Conference (ISWC), 2015, pp.205-221

[2] S. Egami, T. Kawamura, Y. Sei, Y. Tahara, A. Ohsuga, "Desiging and Publishing Illegally Parked Bicycle LOD" International Journal of Smart Computing and Artificial Intelligence, Vol.1, No.2, 2017, pp.77-93

[3] S. Egami, T. Kawamura, A. Ohsuga, "Building Urban LOD for Solving Illegally Parked Bicycles in Tokyo," In: Proceedings of the 15th InternationalSemantic Web Conference (ISWC), 2016, pp.291-307

[4] S. Shiramatsu, T. Tossavainen, T. Ozono, T. Shintani, "Towards Continuous Collaboration on Civic Tech Projects: Use Cases of a Goal Sharing System Based on Linked Open Data," In: Proceeding of the 7th IFIP International Conference on Electronic Participation (ePart), 2015, pp.81-92

[5] H. Santos, V. Dantas, V. Furtado, P. Pinheiro, D, D. L. McGuinness, "From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards," In: Proceeding of the 14th Extended Semantic Web Conference (ESWC), 2017, pp.94-108

[6] S. F. Pileggi, J. Hunter, "An ontological approach to dynamic fine-grained Urban Indicators," Procedia Computer Science, Vol. 108, 2017, pp.2059-2068

[7] K. Höffner, M. Martin, J. Lehmann, "LinkedSpending: OpenSpending becomes Linked Open Data," Semantic Web Journal, Vol.7, No.1, 2016, pp.95-104

[8] G. Demartini, D. E. Difallah, P. Cudré-Mauroux, "Large-scale linked data integration using probabilistic reasoning and crowdsourcing," The International Journal on Very Large Data Bases, Vol. 22, No. 5, 2013, pp.665-687

[9] I. Celino, S. Contessa, M. Corubolo, D. Dell ' Aglio, E. D. Valle, S. Fumeo, T. Krüger, "Linking Smart Cities Datasets with Human Computation - The Case of UrbanMatch," In: Proceedings of the 11th International Semantic Web Conference (ISWC), 2011, pp.34-49

[10] L. V. Ahn, "Games with a purpose," IEEE Computer, Vol. 39, No. 6, 2006, pp. 92-94

[11] T. Kudo, Y. Matsumoto, "Japanese Dependency Analyisis using Cascaded Chunking," In: Proceedings of the 6th Conference on Natural Language Learning, Vol.20, 2002, pp.1-7

[12] W. Winkler, "The state record linkage and current research problems," Technical report, Statistics of Income Division, Internal Revenue Service Publication,1999

[13] T. M. Nguyen, T. Kawamura, Y. Tahara, A. Ohsuga, "Self-Supervised Capturing of Users' Activities from Weblogs," International Journal of Intelligent Information and Database Systems, Vol. 6, No. 1, 2012, pp.61-76

[14] I. Augenstein, S. Padó, S. Rudolph, "LODifier: Generating Linked Data from Unstructured Text," In: Proceedings of the Extended Semantic Web Conference (ESWC), 2012, pp.210-224

[15] D. Milne, I. H. Witten, "Learning to Link with Wikipedia," In: Proceedings of the 17th ACM conference on Information and knowledge management (CIKM), 2008, pp. 509-518

[16] H. Sakaji, S. Sekine, S. Masuyama, "Extracting causal knowledge using clue phrases and syntactic patterns," In: Proceedings of the International Conference on Practical Aspects of Knowledge Management (PAKDD), 2008, pp.111-122

[17] H. Ishii, Q. Ma, M. Yoshikawa, "Causal Network Construction to Support Understanding of News," In: Proceedings of the 43rd Hawaii International Conference of System Sciences (HICSS), 2010, pp.1-10

[18] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, A. Zaveri, "Test-driven Evaluation of Linked Data Quality," In: Proceedings of the 23rd International Conference on World Wide Web (WWW), 2014, pp.747-758

[19] J. L. Fleiss, J. Cohen, "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," Educational and Psychological Measurement, Vol.33, No.3, 1973, pp.613-619

[20] A. J. Viera, J. M. Garrett, ”Understanding interobserver agreement: the kappa statistic,” Fam Med, Vol.37, No.5, 2005, pp.360-363

[21] M. Ashikawa, T. Kawamura, A. Ohsuga, ”Crowdsourcing worker development based on probabilistic task network,” Proceedings of the International Conference on Web Intelligence (WI), 2017, pp.855-862