STI 2018 Leiden

23rd International Conference on Science and Technology Indicators

"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Content-based Map of Science using Cross-lingual Document Embedding – A Comparison of US-Japan Funded Projects

Takahiro Kawamura*, Katsutaro Watanabe*, Shusaku Egami*, Naoya Matsumoto* and Mari Jibu*

**takahiro.kawamura@jst.go.jp*
Japan Science and Technology Agency, Tokyo, 102-8666 (Japan)

Introduction

Since Price (1965) proposed using scientific methods to study science in 1965, research in scientometrics has developed techniques for analyzing research activities and measuring their relationships, and maps of science were constructed for understanding the structure and spread of science and the interconnection of disciplines. Science and technology enterprises can use the maps of science to anticipate changes, especially those initiated in their immediate vicinity. Research laboratories and universities that are organized according to the established standards of disciplinary departments can understand their environmental changes. Furthermore, the maps are important for policy analysts and funding agencies because research funding is based on quantitative and qualitative scientific metrics.

However, as it is difficult to apply inter-citation and co-citation analysis to ongoing projects and recently-published papers that have inadequate citations and references, we developed a content-based map (Kawamura et al., 2017a, 2017b, 2018), which converts text information, such as funding project descriptions and paper abstracts, into multi-dimensional vectors and calculates content similarities, that is, distances between the vectors. However, comparing content-based maps in different languages remains problematic. Therefore, this paper proposes a method for locating multi-dimensional vectors from English and Japanese documents in the same space by converting sentences to graph structures representing semantic roles.

The remainder of this paper is organized as follows. Section 2 discusses related work, and Section 3 describes our proposed method for creating multi-dimensional vectors from cross-lingual documents. We then evaluate the matching result of the vectors using 1,000 bilingual IEEE papers. Section 4 introduces a map that is created from approximately 34,000 US and Japan funded projects of the National Science Foundation (NSF) and the Japan Society for the Promotion of Science (JSPS) between 2012 and 2015, and we present some findings regarding the national funding trends. Finally, Section 5 provides our conclusions and suggestions for future work.

Related Work

Compared with cross-lingual papers and projects, the simple approach is to use codes in a universal coding system or classes in ontology. However, as funding agencies and publishers generally use their own classification systems, no comprehensive scheme for characterizing projects or articles exists, thus making it difficult to make direct comparisons between different agencies or publishers. For example, comparing articles from the Association for

Computing Machinery classification (https://www.acm.org/publications/class-2012) with the Springer Nature classification requires taxonomy exchanges. Archambault et al. (2011) proposed the Open Scientific Journal Ontology for comparing multi-lingual papers, but it has not yet been widely used.

Therefore, we considered calculating content similarities from text information and making clusters of similar documents. Since the meaning of a word is determined by its context (Firth, 1957), document embedding (Le & Mikolov, 2014), which represents the features of words appearing around target words with the word orders, is considered to be more accurate than conventional bag-of-words approaches, such as co-occurrence word analysis and TFIDF. In fact, several studies have conducted bilingual distributed representations (Berard et al., 2016), (Luong et al., 2015), (Gouws et al., 2015).

There are two main approaches to generating bilingual distributed representations. One trains both language models independently and then learns a mapping from one representation to the other, and the other performs the training jointly using a parallel corpus. An advantage of the former bilingual mapping, where vectors are first trained in each language independently and a mapping is learned to transform representations from one language into another, is the training speed, since no further training is required if monolingual vectors are given. In the latter approach, the bilingual training attempts to learn representations jointly from scratch to generate good vectors for both languages. For each pair of sentences in a parallel corpus, bilingual vectors attempt to predict words in the same sentence but they also use words in the source sentence to predict words in the target sentence (and vice versa) (Luong et al., 2015), (Berard et al., 2016). Thus, for each update, the vectors perform four updates: source to source, source to target, target to target, and target to source.

While the accuracy of predicting similar words between different languages still remains less than 50%, converting bilingual word vectors to bilingual document vectors involves simple combinations. For example, document vectors are computed by doing a weighted sum of word vectors, according to the word frequencies. As the result, there is currently no standard method for constructing cross-lingual document vectors.

Apart from word and document embedding techniques, a report of the semantic evaluation challenge (SemEval) (Cer et al., 2017) demonstrates a wide variety of methods for measuring multilingual textual similarity. In a task for semantic textual similarity for multilingual and cross-lingual focused evaluation, the gradation of meaning overlaps is measured between cross-lingual pairs of English with materials in Arabic, Spanish, and Turkish. Since there is no result between English and Japanese, the presented methods and results cannot be directly compared with our proposed method; however, the top four systems all used machine translation (MT) first and then extracted several features for machine learning, such as n-grams, edit distance, and longest common substring. Since 2016, deep learning methods have also been combined with MT. Therefore, in the next section, we compare our cross-lingual document embedding technique from semantic role graphs with machine-translated monolingual document embedding and evaluate their matching accuracy.

Cross-lingual document vectors from SR Graph

This section proposes a method for generating multi-dimensional vectors from documents written in different languages. The simple way to generate cross-lingual vectors is to use MT, which translates one language into another and applies monolingual document embedding. However, this method depends heavily on the accuracy of an MT engine dedicated for each combination of source and target languages. Moreover, statistical MT engines such as Google Translate that are currently the mainstream often change the sentence structure, e.g., from active to passive, and amend phrases to make natural and fluent sentences. Thus, to avoid this influence of the MT engines, this paper first constructs semantic role (SR) graphs from

English and Japanese sentences, and then generates document vectors from both SR graphs, based on a triplification technique (Kawamura et al., 2016) that is often used for generating Resource Description Framework (RDF, https://www.w3.org/RDF/) data from natural sentences. The overall flow consists of the three steps described in the following sections.

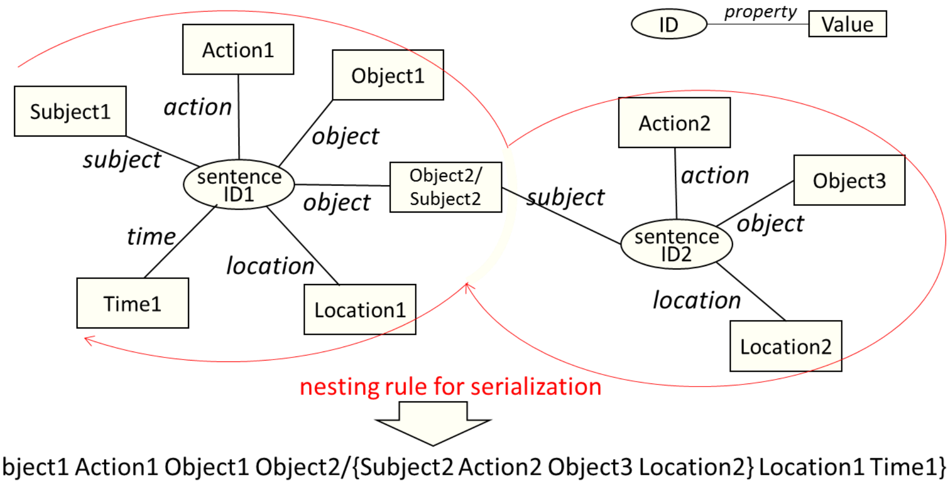
Converting to semantic role graph

First, semantic role labeling (SRL) in natural language processing (NLP) converts sentences written in English and Japanese to SR graphs. SRL is one of the research topics aiming to raise the accuracy of NLP applications, such as MT and question answering, and to extract semantic relations between or among words in natural sentences, such as Who (Subject, Agent), What (Object), Whom (Patient), and How (Action, Predicate). In case grammar (Fillmore, 1968), a semantic structure of sentences is a set of "verb – deep case – noun" relations, and the relations extracted by the SRL correspond to the deep cases. However, the target relations to be extracted vary in research (Mooney , 2014), (Sammons, 2014a).

In this paper, the relations to be extracted are limited to subjects, actions, and objects, according to Sammons (2014b). We also extracted complements as objects to incorporate them into graphs, and location and time. If there are multiple subjects and objects in a sentence, triples (sentence ID, property, and value) for each subject and object are generated. A complex sentence is converted to multiple triples with different IDs. A sentence with an adjectival verb, such as an attributive modification clause, is also divided into the main clause, and the subordinate clause and is converted to multiple triples. In this case, if the subordinate clause has both a subject and object, the triple has a different sentence ID. If the clause has either a subject or object, the triple has a link to the subject of the main clause and the same ID. We also generate triples with the passive verbs as values of action properties.

In terms of restrictions, co-reference relations using indication words between sentences are not resolved. In a sentence, however, demonstrative pronouns are replaced with the preceding subjects or objects, although our survey showed that scientific documents contain few indication words. For the same reason, zero anaphoric relations between sentences, or rather, omissions of the corresponding cases, are not resolved. However, if there is a preceding subject in the same sentence, it replaces the omitted noun. As a result, the sentences become like the skeletons shown in Fig. 1.

Figure 1: Semantic role graph and serialiazation.



Unifying words to descriptors in the scientific thesaurus

Next, after lemmatizing each term in the graph, if the terms are matched to synonyms and/or descriptors in the Japan Science and Technology Agency (JST) science and technology thesaurus and large dictionary (hereafter, JST thesaurus) (Kimura et al., 2015), both English and Japanese terms are replaced with English descriptors in the JST thesaurus. For example, English synonyms for artificial intelligence, such as AI and computational intelligence, and Japanese ones, such as 人工知能 and 計算知能 are all replaced with the descriptor "artificial intelligence."

The JST thesaurus primarily consists of keywords that have been frequently indexed in 36 million articles accumulated by the JST since 1975. The thesaurus is updated quarterly and includes 276,179 terms in English and Japanese from 14 categories ranging from bioscience to computer science and civil engineering. Based on the World Wide Web Consortium Simple Knowledge Organization System (skos), the JST thesaurus exists in RDF format with relationships *skos:broader*, *skos:narrower*, and *skos:related*. A broader or narrower relationship essentially represents an *is-a* subsumption relationship but sometimes denotes a *part-of* relationship in geography, body organ terminology, and other academic disciplines. The JST thesaurus is publicly accessible from Web APIs on the J-GLOBAL website (http://jglobal.jst.go.jp/en/), along with the visualization tool Thesaurus Map (http://thesaurus-map.jst.go.jp/jisho/fullIF/index.html).

In this step, words that are not included in synonyms and descriptors are deleted as non-technical terms, excluding named entities and numerical values.

Generating document vectors

Finally, the graphs are serialized with a simple nesting rule that iteratively lists the values in the same order as shown in Fig. 1. The document vectors are generated from the word sequences.

A word vector is represented as a matrix whose elements are in principle the co-occurrence frequencies between a word *w* with a certain usage frequency in the corpus and words within a fixed window size *c* from *w*. A popular representation of word vectors is word2vec (Mikolov et al., 2013a, 2013b). Word2vec generates word vectors using a two-layered neural network obtained by a skip-gram (or, continuous bag of words) model. Specifically, word vectors are obtained by calculating the maximum likelihood of objective function *L* in Eq. (1), where *T* is the number of words with a certain usage frequency in the corpus. Word2vec clusters words with similar meanings in a vector space.

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \qquad (1)$$

Additionally, Le & Mikolov (2014) proposed a document vector that learns fixed-length feature representations using a two-layered neural network from variable-length pieces of texts such as sentences, paragraphs, and documents. A document vector is considered another word in a document and is shared across all contexts generated from the same document but not across documents. The contexts are fixed length and sampled from a sliding window over the document. The document vectors are computed by fixing the word vectors and training the new document vector until convergence, as shown in Eq. (2).

$$L = \sum_{t=1}^{T}\log p(w_t|w_{t-c}, ..., w_{t+c}, d_i) \qquad (2)$$

where d_i is a vector for a document *i* that includes w_t. Whereas word vectors are shared across documents, document vectors are unique among the documents and represent the topics of the documents. By considering word order, document vectors also address the weaknesses of the bag-of-words approaches and are therefore considered more accurate representations of the context of the content.

We calculate the information entropy of each concept in the JST thesaurus from the dataset. Shannon's entropy (1948) in information theory is an estimate of event informativeness. We used this entropy to measure the semantic diversity of a concept in a vector space. After creating clusters according to the degree of entropy, we unified all word vectors in the same cluster to a cluster vector and constructed document vectors based on the cluster vectors. Using high-entropy concepts, which are significant in scientific and technological contexts as elements between paragraph vectors, the paragraph vectors can comprise meaningful clusters. Previous literature provides more details (Kawamura et al., 2017a, 2017b, 2018).

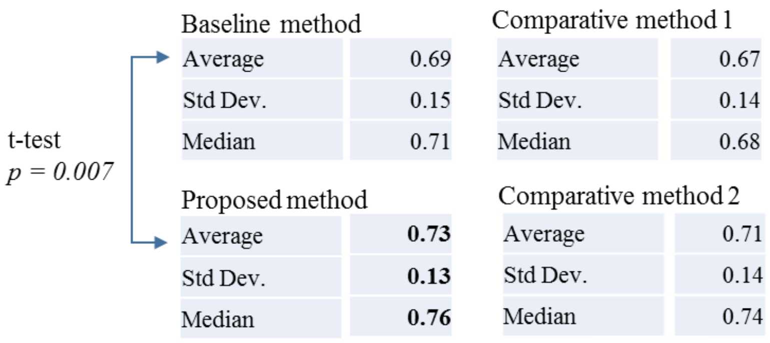
Accuracy of cross-lingual document vectors
The JST provides a bibliographic database, J-GLOBAL (http://jglobal.jst.go.jp/en/), in which titles and abstracts of papers in English are translated into Japanese. The translations are done by translators of scientific and technical literature, who have been involved in building bibliographic databases for many years. Thus, as an experimental dataset, we randomly selected 1,000 paper titles and abstracts in English and the corresponding 1,000 Japanese titles and abstracts from approximately 63,000 IEEE journal and conference papers published between 2012 and 2015. SR graphs were converted from 2,000 titles and abstracts in English and Japanese using TEXT2LOD (Kawamura et al., 2016) based on conditional random fields (Lafferty et al., 2001). The vector space was generated using the method presented in the previous section. Hereafter, paper abstracts refer to the titles and abstracts of papers, and project descriptions include their titles. We evaluated these cross-lingual document vectors based on the following two aspects:

1. The similarity between a document vector $\boldsymbol{ve}_i$ from an English abstract and a document vector $\boldsymbol{vj}_i$ from a Japanese abstract translated by experts.
2. The correlation of the similarity between vectors $\boldsymbol{ve}_i$ and $\boldsymbol{ve}_j$ from two English abstracts and the similarity between vectors $\boldsymbol{vj}_i$ and $\boldsymbol{vj}_j$ from the corresponding two Japanese abstracts translated by experts.

We implemented the document embedding technique using the Deep Learning Library for Java (https://deeplearning4j.org). Despite needing a more systematic method, the hyperparameters were set empirically as follows: 500 dimensions were established for words that appeared more than five times; the window size c was 10, and the learning rate and minimum learning rate were 0.025 and 0.0001, respectively, with an adaptive gradient algorithm. The learning model is a distributed memory model with hierarchical softmax.
The similarity was measured using the cosine similarity of two vectors as well as the SemEval challenges. The baseline method to compare involved document embedding directly generated from English abstracts and those generated from English sentences translated from Japanese abstracts by Google Translate.

1. Comparison of vectors from English and Japanese abstracts of the same document
Table 1 shows the cosine similarities of the average, median, and standard deviation between vectors from the English and Japanese abstracts of the same papers. The comparative method 1 involves adding unification to descriptors and document embedding with entropy clustering to the baseline, although word orders are decided by the MT engine as well as the baseline. The comparative method 2 involves subtracting the conversion to SR graphs from the proposed one so that word orders are the same as the original English and Japanese sentences.

Table 1. Cosine similarities of the average, median, and standard deviation.



Baseline method		Comparative method 1	
Average	0.69	Average	0.67
Std Dev.	0.15	Std Dev.	0.14
Median	0.71	Median	0.68
Proposed method		**Comparative method 2**	
Average	**0.73**	Average	0.71
Std Dev.	**0.13**	Std Dev.	0.14
Median	**0.76**	Median	0.74

t-test
p = 0.007

Consequently, the proposed method indicated the statistically significant difference from the baseline. By comparing the baseline with the comparative method 1, embedding from SR graphs is considered to increase the matching accuracy. This finding is also supported by comparisons with the proposed method and the comparative method 2. The difference between the proposed method and the comparative method 2 was relatively small because unification of words to descriptors shortens the sentence length. By contrast, unlike other methods that eliminate non-technical words and reduce notation variability by unifying words to descriptors in the thesaurus, the baseline showed the bad result since the English translation using the MT of Japanese translation by experts had little reversibility, and the non-technical words and alternative notations of the same meaning in abstracts affected the cosine similarity. In principle, the cosine similarity should be 1.0, but in this dataset, the Japanese translation is free instead of being word-for-word due to copyright issues. Thus, the similarity values were inevitably less than 1.0. Moreover, since a few Japanese abstracts were largely rewritten from the original English abstracts, the median presented the overview of the results well in the sense that it eliminated any outliers. Although the conversion accuracy of SR graphs depends on semantic roles, such as subjects and actions, and varies from 67% to 98%, the weighted average was 94% (Kawamura et al., 2016).

2. Comparison of similarity between English abstracts and between Japanese abstracts

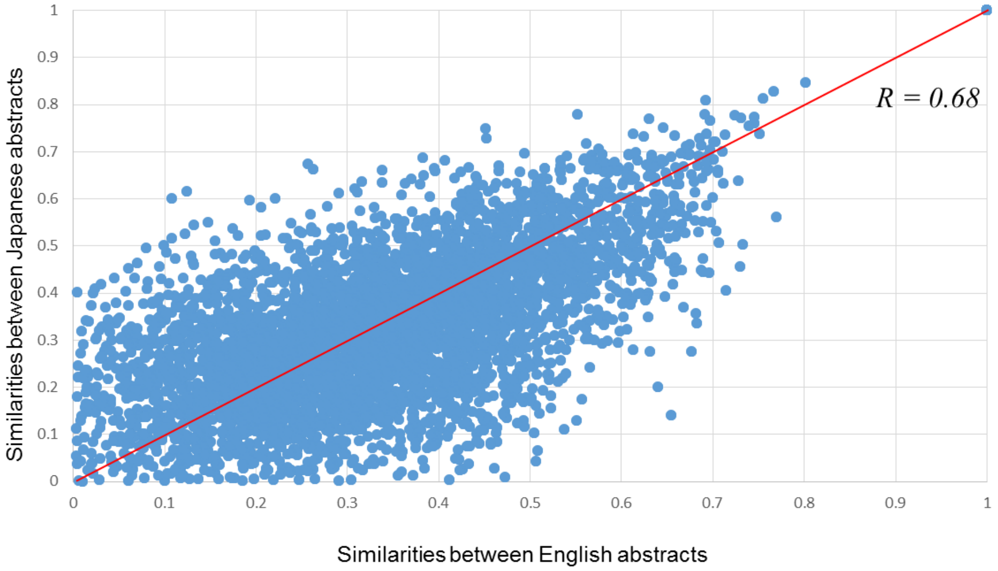
For all pairs of papers in the dataset, Fig. 2 shows the relationship between the similarity of vectors $\boldsymbol{ve}_i$ and $\boldsymbol{ve}_i$ from English abstracts and the similarity of vectors $\boldsymbol{vj}_i$ and $\boldsymbol{vj}_j$ from the corresponding Japanese abstracts.

Consequently, the following correlation

$$R \;=\; correl\;(\;cos(\boldsymbol{ve}_i, \boldsymbol{ve}_j),\; cos(\boldsymbol{vj}_i, \boldsymbol{vj}_j)\;) \qquad (3)$$

was 0.68 in all cosine ranges, as shown in Fig. 4. In practice, however, low similarities are ignored at the map layout phase and are not important; thus, if the similarities are limited to more than 0.5 then *R* became 0.77 and indicated a high correlation. By contrast, the baseline indicated *R* = 0.70.

Figure 2: Relationship between the similarity of vectors from English abstracts and the similarity of vectors from the corresponding Japanese abstracts.



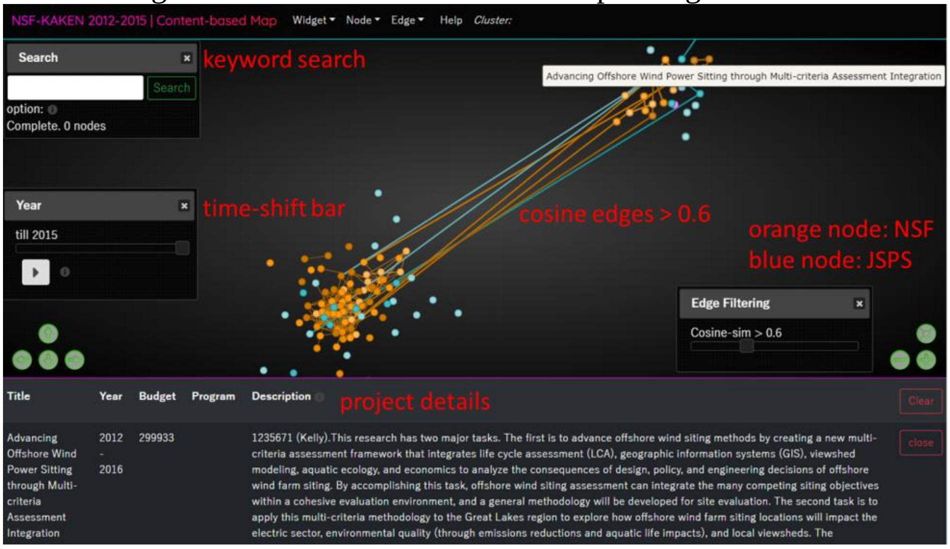
US-Japan Funded Projects Map

On the maps shown in Fig. 3 and 4, the dataset contains titles and descriptions of 25,758 NSF projects (https://federalreporter.nih.gov/) from 2012 to 2015, including 524,509 sentences and those of 8,643 JSPS projects (https://kaken.nii.ac.jp/en/) for the same period, including 101,784 sentences. NSF project domains are limited to Computer & Information Science & Engineering, Mathematical & Physical Sciences, and Engineering, and JSPS project domains are limited to Informatics and Engineering.

Although there are differences between paper abstracts and project descriptions, e.g., descriptions are more formally written than abstracts, we found no critical difference that affects the presented method. We thus generated vectors from the above dataset using the presented method and calculated the cosine similarities of all pairs. We then performed the community detection optimizing modularity scores on edges that indicate more than 0.4 cosine similarities. The threshold 0.4 was determined empirically through the experiments. This time there were 1,834 nodes (projects) that were out of any communities. We used the same tool presented in the previous literature (Kawamura et al., 2017a, 2017b, 2018) for displaying the map, in which distances between the nodes are proportional to the cosine similarities as much as possible. In addition, the map provides functions for searching titles and descriptions by keywords, showing the project details, including titles, descriptions, organizations, years, and budgets, displaying the cosine values on the edges, and querying the graph using W3C SPARQL.

The followings are some findings obtained from the map. Figure 3 presents two communities for the wind-power generation, in which communities are clearly separated for the offshore power generation and controls with safety issues, while NSF and JSPS projects are mixed in both communities. Despite that the number of NSF projects is three times more than the number of JSPS projects, these are almost even in the communities; thus, we found that Japan is working hard on this topic.

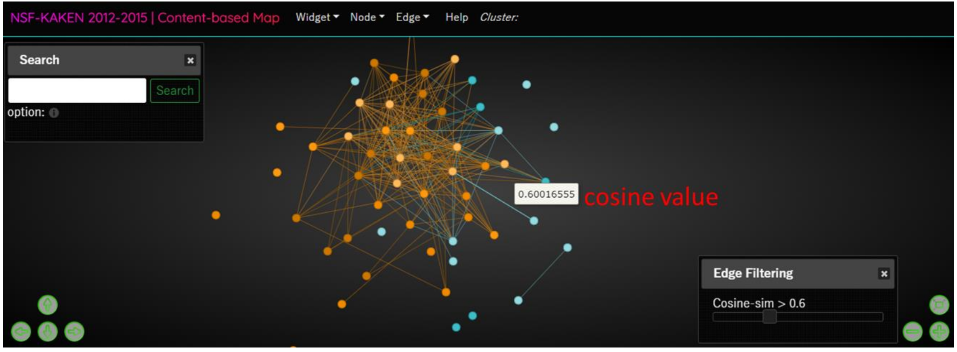
Figure 3: Communities for the wind-power generation.



Title	Year	Budget	Program	Description
Advancing Offshore Wind Power Sitting through Multi-criteria Assessment Integration	2012 - 2016	299933		1235671 (Kelly).This research has two major tasks. The first is to advance offshore wind siting methods by creating a new multi-criteria assessment framework that integrates life cycle assessment (LCA), geographic information systems (GIS), viewshed modeling, aquatic ecology, and economics to analyze the consequences of design, policy, and engineering decisions of offshore wind farm siting. By accomplishing this task, offshore wind siting assessment can integrate the many competing siting objectives within a cohesive evaluation environment, and a general methodology will be developed for site evaluation. The second task is to apply this multi-criteria methodology to the Great Lakes region to explore how offshore wind farm siting locations will impact the electric sector, environmental quality (through emissions reductions and aquatic life impacts), and local viewsheds. The

Figure 4 presents a community for the Terahertz radiation, in which NSF projects and JSPS projects are separately located on the left and right sides. In this topic, Japan puts effort into the application to manufacturers, such as nondestructive inspection, and the US has many applications for wireless communications. Additionally, in a community of unmanned aerial vehicles, Japan seems to have many applications for resource mapping, while data collection when the forest fire happened is attracting attention in the US.

Figure 4: Community for the Terahertz radiation.



In terms of the current limitation, synonyms, such as CS for computer science, Cesium, and consumer satisfaction happen to increase the similarities of unrelated pairs. Due to the graph layout algorithm, when any pair of different topics becomes closer, the whole communities that correspond to these topics are also located close on the map. In future, we plan to replace acronyms with full words before making vectors. The maps will be publicly available at https://jipsti.jst.go.jp/foresight/content-based_map/.

Conclusion and Future Work

In an attempt to resolve the difficulty of a content-based map to compare documents in different languages, this paper proposed a method for generating multi-dimensional vectors in the same space from cross-lingual (English and Japanese) papers/projects. We confirmed a similarity of 0.76 for matching the bilingual contents of 1,000 IEEE papers. Finally, we constructed a map of 34,000 NSF and JSPS projects from 2012 to 2015. As described in the introduction, research evaluators, scientific policy-makers, and funding providers can investigate this map for finding common or different points of interest and the weakness of a nation, comparing national funding trends.

Our next step is to compare citation-based methods and incorporate patent information into the map. In addition, we aim to extract metrics from chronological changes of the network structure in the map. As the Foresight and Understand from Scientific Exposition (FUSE) program in Intelligence Advance Research Projects Activity (IAPRA) conducted a study to identify emerging research areas based on several metrics obtained from several maps of science from 2011 to 2015, we, the Japan Science and Technology Agency, will also apply these metrics to statistical analysis and machine learning techniques in an attempt to detect emerging research areas in their early stages.

References

Archambault, E., Beauchesne, O.H. & Caruso, J. (2011). Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology, *Proceedings of the 13th International Conference on Scientometrics & Informetrics* (ISSI 2011), 66-77.

Berard, A., Servan, C., Pietquin, O. & Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP, *Proceedings of the 10th edition of the Language Resources and Evaluation Conference* (LREC 2016).

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, *Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval-2017), 1-14.

Fillmore, C.J. (1968). The Case for Case, *Universals in Linguistic Theory*, 1-88.

Firth, J.R. (1957). A Synopsis of Linguistic Theory 1930-1955, *Studies in Linguistic Analysis*, 1952-59, 1-32.

Gouws, S., Bengio, Y. & Corrado, G. (2015). BilBOWA: Fast Bilingual Distributed Representations withoutWord Alignments, *Proceedings of the 32nd International Conference on Machine Learning* (ICML 2015), 748-756.

Kawamura, T. & Ohsuga, A. (2016). Development of Web API for Triplification of Text Information, *New Generation Computing*, 34(4), 307-321.

Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S. & Jibu, M. (2017a). Funding Map for Research Project Relationships using Paragraph Vectors, *Proceedings of the 16th International Conference on Scientometrics & Informetrics* (ISSI 2017), 1121-1131.

Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S. & Jibu, M. (2017b). Science Graph for characterizing the recent scientific landscape using Paragraph Vectors, *Proceedings of the 9th ACM International Conference on Knowledge Capture* (K-Cap 2017), 9-16.

Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S. & Jibu, M. (2018). Funding map using paragraph embedding based on semantic diversity, *Scientometrics*, 10.1007/s11192-018-2783-x.

Kimura, T., Kawamura, T., Watanabe, K., Matsumoto, N., Sato, T., Kushida, T. & Matsumura, K. (2015). J-GLOBAL knowledge: Japan's Largest Linked Data for Science and Technology, *Proceedings of the 14th International Semantic Web Conference* (ISWC 2015).

Lafferty, J.D., McCallum, A. & Pereira, F.C.N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the 18th International Conference on Machine Learning* (ICML 2001), 282-289.

Le, Q. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on Machine Learning* (ICML 2014), 32(2), 1188-1196.

Luong, M.T., Pham, H. & Manning, C.D. (2015). Bilingual word representations with monolingual quality in mind, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (NAACL-HLT 2015), 151-159.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at the International Conference on Learning Representations* (ICLR 2013).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems* (NIPS 13), 2, 3111-3119.

Mooney, R.J. (2014). Natural Language Processing, Retrieved March 31, 2018 from: http://www.cs.utexas.edu/~mooney/cs388/slides/srl.ppt.

Price, D. (1965). Networks of Scientific Papers. *Science*, 149, 510-515

Sammons, M. (2014a). Semantic Parsing (Semantic Role Labeling), Retrieved March 31, 2018 from: http://cogcomp.cs.illinois.edu/page/project_view/7.

Sammons, M. (2014b). Semantic Role Labeling Demo, Retrieved March 31, 2018 from: http://cogcomp.cs.illinois.edu/page/demo_view/srl.

Shannon, C. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, 379-423, 623-656.