WISDOM-DX: An Automatic DX Evaluation System Using a QA System Based on Web Information

AKITOSHI OKUMURA[1]　KAI ISHIKAWA[1]　DAI KUSUI[1]
NORIYOSHI ICHINOSE[1]　KENTARO TORISAWA[2]　KIYONORI OHTAKE[2]

Abstract: The promotion of digital transformation (DX) is an urgent issue for Japanese society. To promote companies' DX initiatives, various surveys on DX have been manually conducted by private research companies, industry associations, local governments, and government agencies. However, 95% of companies are either not working on DX at all or are only in the beginning stage of working on it. They have difficulty understanding the purpose and methods of DX that are appropriate for them. Although the surveys introduce the general DX trends and the DX initiatives of top-ranked companies, it is difficult for most of the companies to recognize their own positions and to find referable good practices from the surveys. Although it is necessary and effective for the companies to make objective evaluations for benchmarking such as scoring their DX initiatives and rankings among other companies, it is not easy for them to conduct the benchmark surveys themselves, which require designing the evaluation items, conducting the evaluation, and benchmarking for DX promotion, because the survey cost in time and expense is not small. Instead of this kind of manual survey, Web information could be helpful in conjunction with a sophisticated search technique because companies that are active in DX disseminate a lot of information on the Web through public relations, investor relations, and other promotional activities. However, it has not been clarified what kind of queries are effective for benchmarking DX initiatives. There is no reported method for obtaining the appropriate Web information of companies' DX and evaluating the companies using the information. To make it possible for companies to make objective evaluations, this paper proposes WISDOM-DX, a system that leverages a question answering (QA) system based on Web information that automatically evaluates companies' DX initiatives. By modeling evaluation items in the form of 5W1H (when, who, where, what, why, how) questions, WISDOM-DX evaluates DX initiatives by scoring an answer set generated by the QA system. WISDOM-DX thus makes it possible to obtain consistent benchmark results in a timely, efficient manner. To examine the feasibility of using Web data, WISDOM-DX and a baseline method that used Google Custom Search were evaluated by ranking 464 companies that responded to the DX Stocks 2021 survey from which DX experts selected 48 companies for distinction as DX Stocks 2021 or Noteworthy DX Companies 2021. Regarding the top 48 companies ranked by WISDOM-DX, 27 of them were included among the 48 selected companies and 17 of them had received DX-related awards or certifications, indicating that 91.7% had a certain level of achievement for their DX initiatives. In contrast, 11 of the top 48 companies ranked by the baseline method were included among the 48 selected companies and 20 of them had received DX-related awards or certifications, indicating that 64.6% had a certain level of achievement for their DX initiatives. When WISDOM-DX and the baseline method were evaluated for searching for the 48 selected companies, the area under the precision-recall curve (AUPR) values obtained by WISDOM-DX and the baseline method were 0.541 and 0.181, respectively. In addition, the respective precision values were 56.3% and 22.9%. The survey of WISDOM-DX with the questionnaire to the evaluated companies showed that 60.7% offered positive responses and 32.1% neutral responses regarding the agreeability of their rankings, and that 46.4% offered positive responses and 39.3% neutral responses regarding the usefulness of the system. These results show that WISDOM-DX had more promising performance than the baseline method, and that it offers the prospect of automating large-scale analysis and evaluation of DX initiatives as a first step in using Web data for benchmarking companies. We will provide support functions to improve WISDOM-DX for practical use by companies and research organizations.

1. Introduction

The Japanese government has been promoting the digital transformation (DX) of society as a whole from the perspective of data utilization and digital government, with a view toward achieving the goals of the Society 5.0 initiative [1,2]. In response to these efforts, government agencies are required to develop DX promotion plans and evaluate the results promptly by using evidence [3]. Priority plans for the formation of a digital society, approved by the Cabinet in December 2021, addressed the importance of obtaining evidence through continuous, real-time data acquisition [2]. In other words, DX promotion requires evidence-based evaluation in a timely manner.

The DX Promotion Guidelines published in 2018 defined DX as the transformation of products, services, and business models according to the needs of customers and society, as well as the transformation of operations, organizations, processes, and culture to establish competitive advantages, by using digital data and technology to adapt to rapid changes in business environments [4]. The Ministry of Economy, Trade and Industry (METI) published the DX Report [5] in September 2018 to promote DX in companies. It has also been developing policies from the inside out and from the outside in by improving market environments. In December 2020, METI published the DX Report 2 (interim report) [6], which outlined the need to break away from legacy enterprise culture and promote co-creation between user and vendor companies. In August 2021, it published the DX Report 2.1 [7], which outlined the shape of industries and companies after DX, as well as issues and policy directions for accelerating the transformation of companies. To promote their DX initiatives, METI has developed DX evaluation schemes. These include the Digital Governance Code, which describes an ideal form of digital governance [8], the DX Certification System [9], and the selection of DX Stocks [10], as well as the DX Promotion Guidelines [4] and the DX Promotion Index [11]. Surveys based on such evaluation schemes can serve as a basis

[1] Information-technology Promotion Agency, Japan
[2] National Institute of Information and Communications Technology

© 2022 Information Processing Society of Japan

for top management to plan and execute DX-related business strategies. They need to address these reasons and understand changing business environments, including markets and competitors, to establish a competitive advantage [13].

However, Japanese companies are still far behind U.S. companies in these regards, although they have been improving their IT infrastructure and employment policies in areas such as telework to cope with the business continuity crisis caused by COVID-19 [12]. Self-diagnosis in terms of the DX Promotion Index [11] showed that 95% of companies are either not working on DX at all or are only in the beginning stage of working on it, which indicates a large difference in the status of DX promotion between leading and average companies [6]. Regarding the reasons for the lack of progress in DX by certain companies, it has been suggested that they do not understand the purpose of DX, what to do for it, and how to proceed with it [13]. They have difficulty understanding the purpose and methods of DX that are appropriate for them because most of the surveys only introduce the general trends of DX and DX initiatives of top-ranked companies. To recognize their positions and to find referable good practices, they have to conduct benchmark surveys for themselves. As illustrated in Figure 1, the surveys require designing evaluation items, conducting the evaluation, and benchmarking for DX promotion.

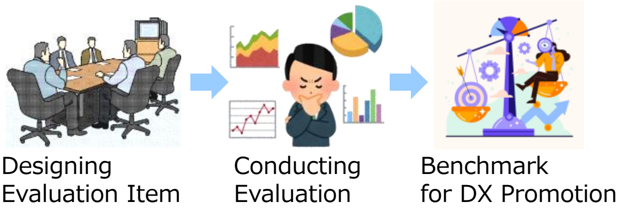


Figure 1: DX Benchmark Survey Process

DX evaluation items constitute questions for target companies, which should be designed to be consistent and objective to capture time-series trends and to enable comparative analysis among companies. These evaluation items are often formulated via questionnaires, and respondents (or investigators) prepare their responses after checking the status of each evaluation item. The responses are then rated by evaluators such as DX experts. The evaluation results are scored for each evaluation item and visualized in the form of a radar chart or an overall ranking. As the numbers of items and respondents increase, the cost in time and expense increases accordingly for the evaluation item designers, the questionnaire respondents, and the evaluators. There are several reasons for the increased time and expense. Especially, the cost for evaluators to rate responses is not low. In particular, analysis of respondents' free statements (qualitative analysis) is not an easy task even for experts. Because qualitative analysis lacks standardized methods like those of quantitative analysis [14], the process of interpretation can be arbitrary and unclear [15]. Problems of oversight and subjective bias have also been pointed out for qualitative analysis, because a single expert can only grasp a limited amount of data [16]. As a result, it is necessary for multiple evaluators to deliberate on the results of each evaluation from various perspectives and compile them into final evaluation results such as rankings. Although it is necessary for top management to promote DX with benchmarking, it is difficult for them to conduct the benchmark surveys themselves because of the cost in time and expense.

Instead of this kind of manual survey, Web information could be helpful in conjunction with a sophisticated search technique because companies that are active in DX disseminate a lot of information on the Web through public relations, investor relations, and other promotional activities. To date, there is no reported method for obtaining the appropriate Web information of companies’ DX and evaluating the companies using the information. It will be necessary to clarify the search and evaluation techniques. To make it possible for companies to make objective evaluations, this paper proposes WISDOM-DX, a system that leverages a question answering (QA) system based on Web information to automatically search for and evaluate companies’ DX initiatives without a manual DX survey. By modeling evaluation items in the form of 5W1H (when, who, where, what, why, how) questions, WISDOM-DX evaluates DX initiatives by scoring an answer set generated by the QA system. It is designed to support the evaluation work of companies and research institutes by automating survey processes that currently depend on human labor. WISDOM-DX supports DX promotion by providing companies with relative rankings of their DX initiatives and making it easier for them to compare themselves with other companies. It can also help in ranking the DX initiatives of a larger number of companies and evaluate them from multiple perspectives, such as industry types and company sizes. We thus aim to achieve evidence-based, large-scale, timely surveys with WISDOM-DX. Accordingly, in this paper, we clarify the feasibility of evaluation as a first step in using Web data for this purpose, by comparing evaluation results from WISDOM-DX with those from expert evaluations of DX initiatives.

The remainder of the paper is organized as follows. In Section 2, we introduce the related work on DX surveys in Japan, text analysis with natural language processing, and QA systems. In Section 3, we describe the system configuration of WISDOM-DX, its generation of 5W1H questions and answer sets, and its scoring. In Section 4, we describe our experiments on applying WISDOM-DX to the task of evaluating 464 companies that responded to the DX Stocks 2021 survey, and we report the results in comparison with the 48 companies that were actually selected by expert evaluators. In Section 5, we discuss the differences in the results between WISDOM-DX and the evaluators. Finally, in Section 6, we outline future applications of WISDOM-DX.

2. Related Work

2.1 DX Surveys in Japan

In recent years, a number of organizations such as private research companies [17,18,19], industry associations [20,21,22], local governments [23,24], and government agencies [10,25,26,27] have conducted surveys to analyze the progress of DX and to review the adoption of grant projects. The surveys have targeted private companies, public institutions such as

© 2022 Information Processing Society of Japan

municipalities and government agencies, and various other kinds of organizations. They are generally evaluated by experts who analyze the results of questionnaires, interviews, and proposals. The answers to questionnaires can be in either a selective or free-text format. The selective format requires respondents to choose an answer from a list of prepared options, whereas the free-text format allows them to answer in their own words. The selective format is easy to analyze quantitatively by securing a large amount of data and classifying the respondents, while the free-text format is more suitable for qualitative analysis of respondents' arguments and intentions.

In the case of the DX Stocks selected by Japan's Ministry of Economy, Trade and Industry (METI) and the Tokyo Stock Exchange (TSE), the Evaluation Committee selects outstanding companies according to the results of questionnaires using both the selective and free-text formats [10,14]. The companies selected as DX Stocks are those that have been recognized not only for introducing outstanding IT systems and using data, but also for continuing to take on the challenge of reforming their business models and management through the application of digital technology. In 2015, to promote strategic IT utilization in Japanese companies, METI and TSE began selecting certain companies as Competitive IT Strategy Company Stocks [10]. Specifically, they selected companies that actively apply IT to facilitate management innovation, raise profit levels, and improve productivity with the goal of enhancing corporate value and competitiveness in the medium and long terms. Since 2020, METI and TSE have selected DX Stocks instead of Competitive IT Strategy Company Stocks [10]. To select the DX Stocks for 2021, the DX Research Secretariat conducted a questionnaire survey of approximately 3,700 companies that were listed on the TSE in November 2020. The survey covered the following six major items: (A) management vision and business model, (B) strategy, (C) use of IT systems and digital technology to implement strategy, (D) organization and scheme to implement strategy, (E) governance, and (F) sharing of results and key performance indicators.

Responses were received from 464 companies that covered the 33 industry types among the TOPIX Sector Indices. In the first step, responding companies were evaluated in terms of their selective answers to 35 questions and their three-year average scores for return on equity (ROE). In the second step, the DX Evaluation Committee, which consisted of nine experts, evaluated the companies' DX initiatives by analyzing their free-text answers to 38 questions. The committee's discussions resulted in the selection of the DX Stocks 2021 (28 companies, including two "Grand Prix" companies) and the Noteworthy DX Companies 2021 (20 companies) in June 2021 [10], for a total of 48 companies that were singled out for distinction. Among the DX Stocks 2021, one or two companies were selected for each of the 33 industry types [25]. The Noteworthy DX Companies 2021 were selected from companies that were not selected among the DX Stocks 2021 but had noteworthy initiatives in the area of corporate value contribution [25]. The overall ranking of the 48 selected companies has not been disclosed. Because the Grand Prix companies were highlighted among the DX Stocks for their particularly outstanding initiatives, the Grand Prix, DX Stocks, and Noteworthy DX Companies were highly evaluated in that order. Note that, from 2015 to 2019, METI and TSE also selected Grand Prix companies, and they selected Noteworthy IT Strategy Companies, before transitioning to the selection of Noteworthy DX Companies in 2020. All the names and initiatives of the selected companies are published on the Web via stock selection reports based on the questionnaire results [10]. The response rate of the questionnaires was between 6% and 15% [26]. By reducing the enormous amount of time and effort of the designers, it would be possible to conduct a large-scale survey in a timely manner.

2.2 Text Analysis with Natural Language Processing

Companies' business strategies are often analyzed through text mining of textual data obtained from free-text responses in questionnaires and interviews [16]. Text mining is a technique for quantitatively analyzing text with qualitative characteristics [15]. The term "text mining" has been in use since the mid-1990s [15], and text mining software packages have been developed for quantitative analysis of text data. This software quantitatively organizes a large amount of linguistic data by extracting words from the data and visualizing relationships among these words. This approach makes it possible to grasp the whole picture of the data, to obtain directions of inquiry, and to mathematically demonstrate the bases of data interpretations [28,29].

By introducing the concept of analogy to corporate strategy cases, Goto et al. presented a form of business strategy analysis that enables managers of small and medium-sized companies to find cases that are most similar to their own situations [30]. In other words, they proposed a model to efficiently find strategy cases that are applicable to a company's business environment. They extracted analogy evaluation indicators that incorporate the perspectives of SWOT analysis, which is a method for corporate strategy planning. From case studies of 202 companies described in articles in the publication *Nikkei Business* [31], 185 analogy evaluation indicators were extracted as sub-items from the five major items of strategy, strengths, weaknesses, opportunities, and threats. As a result, Goto et al. demonstrated that analogy evaluation indicators can be used to efficiently search for strategy cases with a high degree of similarity. Although those results provide useful suggestions for an approach to business analysis, it will be necessary to clarify what kind of search queries should be prepared for obtaining information as well as evaluation indicators for DX when the approach is applied for using Web information.

Automated essay scoring (AES) is a general term for computerized essay scoring tasks, which aim to automatically score answers to written questions [32]. AES systems include e-rater [33], which is used for scoring the TOEFL test, and Jess [34], a short essay evaluation system in Japanese. Research has also been conducted on methods that use support vector machines (SVMs) with bags-of-words as features [35] and neural networks (NNs) to solve text classification problems. NN models include recurrent neural networks (RNNs) [36], which recursively construct sentence vectors based on sentence trees, and convolutional neural networks (CNNs) [37], which can make

© 2022 Information Processing Society of Japan

phrase-by-phrase decisions by convolution of local information. These models have been used in tasks such as polarity judgments for movie reviews and question classification, and they have achieved better accuracy than the SVM baseline method. Terada et al. proposed a method for automatically grading the overall correctness or incorrectness of answers to written questions [38]. They used a CNN to extract useful features in units larger than words and classified the answers; as a result, they achieved 90% accuracy in their experiments with several hundred answers. Their method's effectiveness is unclear for evaluation of statements that do not have self-evident correct answers, such as DX initiative survey responses. Nevertheless, it is a useful reference technique in devising guidelines for scoring DX initiatives.

2.3 Question Answering Systems

QA systems have been studied as a means of automatically generating answers to questions [39,40,41]. Because the answering capability depends on the quality and quantity of the available data, there is an issue of how to obtain and update the data. Web information and Wikipedia have been proposed as knowledge sources [42,43,44]. Large-scale data has been used for answering open questions, and QA techniques have been developed in an international shared task [41]. Task design is a critical issue for practical use because it is not possible to answer all questions completely. WISDOM X is a QA system that uses data from approximately six billion Web pages to answer the following types of questions [44]: the fact-type (e.g., “What will happen with global warming?”, “When did global warming start?”, “Where is global warming occurring?”), how-type (e.g., “How can global warming be prevented?”), why-type [45] (e.g., “Why did global warming worsen?”), what-happens-if-type [46] (e.g., “What happens if global warming worsens?”), and definition-type (e.g., “What is global warming?”). WISDOM X is designed to provide a wide range of pinpoint answers, such as a noun phrase for a fact question or a sentence for a what-happens-if question. This feature constitutes a major difference from commercial search engines, which merely provide Web pages in response to a given question and rely on human effort to ascertain pinpoint answers. WISDOM X has been available since 2015. In March 2021, we improved it by incorporating the BERT model pretrained on 350 GB of text and applying our proprietary technique that combines BERT with a deep learning technique called adversarial learning [46,47,48,49]. This improvement resulted in greater accuracy and increased the variety of questions that can be answered. The improved system was equipped with the middleware RaSC [50] to efficiently run various NLP tools on hundreds of computation nodes. WISDOM X can be licensed for use in system development or database construction with the permission of NICT.

3. WISDOM-DX

3.1 Outline of WISDOM-DX

Excellent DX companies generally make effective use of the Internet and other digital technologies in their business, and their DX initiatives and evaluations are often reported on the Web. In addition to METI's report on the companies selected for DX Stocks [10], various media and research organizations have published excellent corporate initiatives on the Web [51,52,53]. Companies that are active in DX disseminate a lot of information on the Web through public relations, investor relations, and other promotional activities. This includes information such as the direction of corporate management and the use of IT technologies, specific strategies, systems to promote those strategies, measures to improve the business environment, and the status of strategies. There are two types of DX promotion initiatives: those related to corporate management, such as top-management commitment, presentation of management strategy and vision, and organizational development [4]; and those related to IT technologies such as the cloud, Internet of things (IoT), big data, and AI [54]. By using such Web data on corporate management and IT technologies, we have developed a system, called WISDOM-DX, to automatically perform surveys that are currently conducted manually. This system aims at automating each of the processes shown in Figure 1. First, automatic design of evaluation items is implemented by modeling them in the form of 5W1H (when, who, where, what, why, how) questions. Then, answer sets are generated by WISDOM X, which can provide answer passages and plausibility analysis for questions. Finally, WISDOM-DX visualizes evaluation results in the form of an overall ranking by integrating scores that are calculated from the answer volume, answer plausibility, and answer similarity to DX good practices.

Figure 2 shows the system configuration of WISDOM-DX. By composing expressions from a question expansion table and a domain dictionary, the 5W1H question generation module produces a list of 5W1H questions about the DX initiatives of each company in an input company list. Next, the answer set generation module outputs an answer set obtained by inputting the 5W1H question list to WISDOM X. Finally, the scoring module evaluates the answer set from the viewpoints of the answer volume, plausibility, and similarity, and it outputs a company ranking based on DX good practices, training data, and task-dependent rules. The DX good practices consist of text data

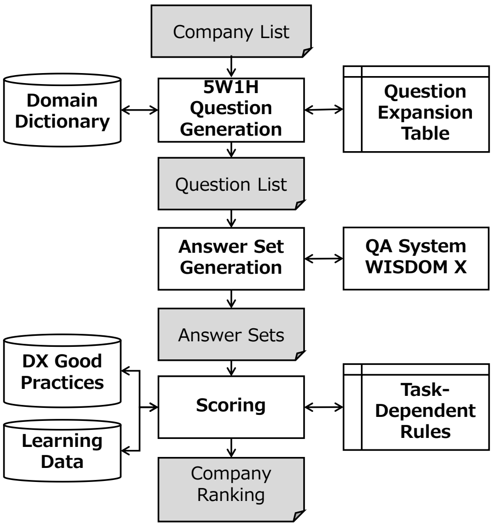


Figure 2: WISDOM-DX System Configuration

© 2022 Information Processing Society of Japan

about company initiatives that have been published on the Web as good practices in the past. The training data is binary and consists of a positive or negative value for each company. Companies that have been reported to have excellent DX initiatives are recorded as positive, while other companies are recorded as negative. Lastly, the task-dependent rules are constraints or conditions that evaluators should consider in addition to the content of initiatives. For example, in the case of DX Stocks 2021, a maximum of one or two companies were selected for each of the 33 industry types to avoid bias toward any particular industry [25]. The task-dependent rules control the final ranking by giving higher priority to the top-level companies in a particular industry.

3.2 5W1H Question Generation

WISDOM-DX generates generic, exhaustive 5W1H questions in accordance with DX evaluation items. The question expansion table contains slots such as <sub>, <obj>, and <pred>. The slots store the following slotted question templates that correspond to each question type:

- Question Type 1: How did <sub> <pred> <obj>?
  (<sub> wa donoyouni <obj> wo <pred> ka?)
- Question Type 2: Where did <sub> <pred> <obj>?
  (<sub> wa dokode <obj> wo <pred> ka?)
- Question Type 3: Who <pred> <obj> in <sub>?
  (<sub> wa dare ga <obj> wo <pred> ka?)
- Question Type 4: What did <sub> <pred> for <obj>?
  (<sub> wa <obj> de nani wo <pred> ka?
- Question Type 5: Why did <sub> <pred> <obj>?
  (<sub> wa naze <obj> wo <pred>ka?)
- Question Type 6: When did <sub> <pred> <obj>?
  (<sub>wa itsu kara <obj> wo <pred> ka?)

For individual question types, the domain dictionary describes specific expressions for each slot, such as “digital transformation” for <obj> and “conduct, achieve” and “start” for <pred>. It also contains company aliases. When the <sub> slot is filled with a company’s name from the input list, the company's aliases are also added to the <sub> slot if the company and its aliases are stored in the domain dictionary. The 5W1H questions are generated using the question type templates and a combination of all the slot expressions. When evaluating the DX initiatives of Company A for the selection of DX Stocks 2021, WISDOM-DX generates the following six types of questions: "How did Company A conduct DX?"; "Where did Company A conduct DX?"; "Who conducted DX for Company A?"; "What did Company A achieve with DX?"; "Why did Company A conduct DX?"; and "When did Company A start DX?"

3.3 Answer Set Generation

The answer set generation module inputs questions one by one to WISDOM X after extracting them from the 5W1H question list. Figure 3 shows the QA model of WISDOM X. After embedding the questions and passages obtained from Web data, WISDOM X inputs them to adversarial networks for generating compact-answer representation (AGR). It also inputs them to a passage encoder and a question encoder, which are BERT-based representation generators. Next, it generates compact-answer representations as fake representations, as well as passage and question representations as true representations. Then, logistic-regression-based answer selection estimates the possibility (plausibility) that each passage contains an answer (positive event) from the true and false representations. Eventually, WISDOM X sends the answer passages with plausibility values to the answer set generation module [47]. For each of the 5W1H question types, the answer set generation module composes an answer triplet consisting of the passages with plausibility values and URLs of Web data. Finally, all of the answer triplets obtained from each question type are merged to form an answer triplet set without duplication. The answer triplet sets use the data structure shown in Figure 4.

3.4 Scoring

3.4.1 Score Functions

As mentioned above, WISDOM-DX scores a company’s DX initiatives in terms of the answer sets and DX good practices from the following viewpoints: answer volume, answer plausibility, and similarity to DX good practices. These characteristics are used to define eight score functions. Specifically, the answer volume is used to define $Score_{cnt}$. Then, the answer volume and similarity to DX good practices are used for $Score_{sim}$, $Score_{sim\_idf}$, and $Score_{sim\_tf\_idf}$. Finally, $Score_{cnt\_p}$, $Score_{sim\_p}$, $Score_{sim\_idf\_p}$, and $Score_{sim\_tf\_idf\_p}$ are derived by respectively combining the plausibility with each of the four previous functions. The eight score functions are formulated with the following notation:

$\boldsymbol{D}$ is an answer set for one of the six question types obtained by WISDOM X.

$\boldsymbol{D}_t$ is a set of answer triplets for question type t.

An element d_t of $\boldsymbol{D}_t$ is a triplet of an answer passage, a URL, and the plausibility obtained by WISDOM X for the answer passage.

$p(d_t)$ is the plausibility of d_t.

w_t is a word contained in the answer passage of d_t, i.e., $w_t \in d_t$.

$\{w_h\}$ is a word set contained in d_h, which is a text consisting of DX good practices.

The answer volume is the total number of elements d_t.

The similarity to DX good practices is represented by the following basic lexical similarity between d_t and d_h.

$$sim(w_t, \{w_h\}) = \max_{\{w_h\}} \frac{\boldsymbol{v}(w_t) \cdot \boldsymbol{v}(w_h)}{\|\boldsymbol{v}(w_t)\| \|\boldsymbol{v}(w_h)\|}$$

Here, $\boldsymbol{v}(w_t)$ and $\boldsymbol{v}(w_h)$ respectively represent the word-embedding vectors of w_t and w_h. They are obtained from morphological analysis with the natural language processing library, spaCy [55], and “ja_core_news_lg,” which is a Japanese language model for spaCy that is derived from UD Japanese GSD [56]. The number of words in this model is 480,000, and the dimension of the vectors is 300. The lexical similarity $sim(w_t, \{w_h\})$ can be enhanced by introducing the term frequency $tf(w_t)$ and the inverse document frequency $idf(w_t)$, which are effective weighting factors used for information

© 2022 Information Processing Society of Japan

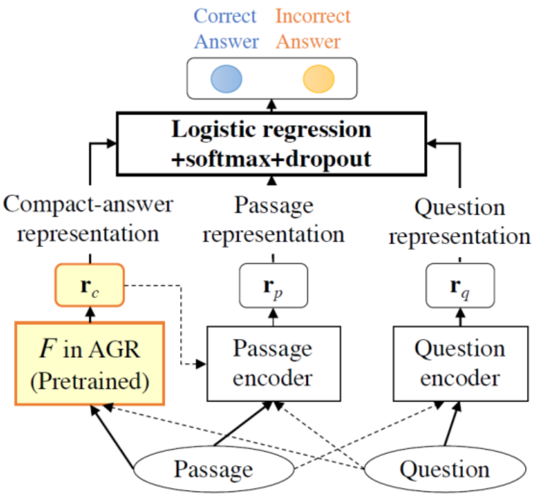
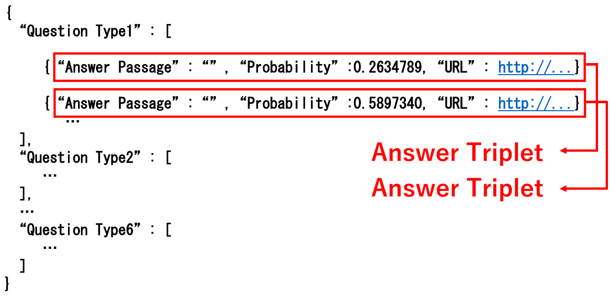


Figure 3: QA Model of WISDOM X
(Excerpt from Figure 1(a) of Reference [47])



```
{
  "Question Type1" : [
     {"Answer Passage" : "" , "Probability" :0.2634789, "URL" : http://...}
     {"Answer Passage" : "" , "Probability" :0.5897340, "URL" : http://...}
     ...
  ],
  "Question Type2" : [
     ...
  ],
  ...
  "Question Type6" : [
     ...
  ]
}
```

Figure 4: Data Structure of the Answer Triplet Set

retrieval.

The eight score functions for the answer triplet set $\boldsymbol{D}_t$ are formulated as indicated below.

Score function ***cnt***: Count of the answer volume.

$$Score_{cnt}(\boldsymbol{D}_t) = \sum_{d_t \in \boldsymbol{D}_t} 1$$

Score function ***sim***: Combination of the basic lexical similarity and ***cnt***.

$$Score_{sim}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\})$$

Score function ***sim_idf***: Combination of the inverse document frequency and ***sim***.

$$Score_{sim\_idf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot idf(w_t)$$

Score function ***sim_tf_idf***: Combination of the term frequency and ***sim_idf***.

$$Score_{sim\_tf\_idf}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot tf(w_t) \cdot idf(w_t)$$

Score function ***cnt_p***: Combination of the plausibility and ***cnt***.

$$Score_{cnt\_conf}(\boldsymbol{D}_t) = \sum_{d_t \in \boldsymbol{D}_t} p(d_t)$$

Score function ***sim_p***: Combination of the plausibility and ***sim***.

$$Score_{sim\_p}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot p(d_t)$$

Score function ***sim_idf_p***: Combination of the inverse document frequency and ***sim_p***.

$$Score_{sim\_idf\_p}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot idf(w_t) \cdot p(d_t)$$

Score function ***sim_tf_idf_p***: Combination of the term frequency and ***sim_idf_p***.

$$Score_{sim\_tf\_idf\_p}(\boldsymbol{D}_t,\ \{w_h\}) = \sum_{d_t \in \boldsymbol{D}_t} \sum_{w_t \in d_t} sim(w_t, \{w_h\}) \cdot tf(w_t) \cdot idf(w_t) \cdot p(d_t)$$

In summary, WISDOM-DX scores answer triplet sets $\boldsymbol{D}_t$ with these eight functions.

3.4.2 Multi-Question Score Ensemble

As described in Section 3.3, WISDOM-DX generates an answer triplet set for each of the six question types. It then applies the eight score functions described in Section 3.4.1 to the six sets, which results in a total of 48 (6x8) scores. To rank companies, it is necessary to obtain an integrated score from the 48 scores. An unsupervised integration method, reciprocal rank fusion (RRF), has been proposed as a way to integrate multiple scores like this. It uses a simple formulation of the reciprocal rank with a constant correction term added without weights. RRF has been reported to obtain better performance than the standard Condorcet integration method and other learning-based methods for integrating multiple relevant document rankings [57] in the NIST TREC document retrieval task [58]. Different combinations of WISDOM-DX's six question types and eight score functions can vary in their accuracy. Weighted averaging, a kind of ensemble method, is known to be advantageous in handling such variation [59]. Given all of these considerations, we developed multi-question score ensemble (MQSE), which is an extended version of RRF that incorporates coupling parameters to obtain integrated scores from the rankings of all question types and score functions.

© 2022 Information Processing Society of Japan

By using training data, MQSE searches for coupling parameters that maximize an evaluation measure for the company ranking, with the integration scores constituting the objective function. When the training data consists of the previous survey results for DX Stocks or Competitive IT Strategy Company Stocks, the label is set to 1 for companies selected as DX Stocks or Noteworthy companies, and to 0 for unselected companies.

The task of evaluating companies on DX initiatives is to obtain a ranking with good accuracy over the entire ranking from top to bottom. Accordingly, the area under the curve (AUC) is preferable as an evaluation metric to the confusion matrix, which fixes the number of positive examples obtained by the system. It is also necessary to account for cases of highly imbalanced data with a low ratio of positive to negative examples. Because the area under the precision-recall curve (AUPR) is more sensitive to the true positive rate (TPR) at the top of the ranking than the area under the receiver operating characteristic curve (AUROC), it is preferable as the objective index. Herein, we use AUC to denote the area under the ROC curve and AUPR to denote the area under the precision-recall curve. The coupling parameters of MQSE are optimized by direct optimization of the final objective index AUPR. Specifically, we use the following procedure to estimate the coupling coefficients.

Step 1: The values of the eight score functions are obtained for the answer triplet set $\boldsymbol{D}_t$ for the six question types and each company.

Step 2: The scores obtained in Step 1 for each company are divided into pairs consisting of each question type and score function, and the scores are then sorted in descending order to obtain a company ranking $rank\left(Score_s(\boldsymbol{D}_t)\right)$.

Step 3: The overall score $Score_{MQSE}$ is obtained from the ranking of the pairs of all question types and score functions in the answer set $\boldsymbol{D}$ by the following formula.

$$Score_{MQSE}(\boldsymbol{D}) = \sum_{\{s,t\}} \frac{\widehat{c_{s,t}}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$$

Here, $\{s\}$ consists of the eight score functions, $\{t\}$ consists of the six question types, and the $\widehat{c_{s,t}}$ are the coupling coefficients.

Step 4: The coupling coefficients are directly optimized by using the AUPR as the objective index, as follows.

$$\widehat{c_{s,t}} = \underset{c_{s,t}}{\operatorname{argmax}} AUPR\left(Score_{MQSE}(\boldsymbol{D}), y_{true}\right)$$

Here, y_{true} denotes the labels for binary classification in the training data. The training data for WISDOM-DX contains the companies selected from 2015 to 2020 as DX Stocks, Competitive IT Strategy Company Stocks, Noteworthy DX Companies, or Noteworthy IT Strategy Companies. The coupling coefficients are estimated by grid search with positive labels for selected companies and negative labels for unselected companies. In addition, $AUPR\left(Score_{MQSE}(\boldsymbol{D}), y_{true}\right)$ is the AUPR of a precision-recall curve obtained by using the ranking results of the score function $Score_{MQSE}(\boldsymbol{D}_t)$ as the argument along with the labels y_{true}.

Although SVM-perf [60] is a direct AUC optimization algorithm and could be a specific method of estimating the $\widehat{c_{s,t}}$, it cannot be applied to optimize the MQSE coupling coefficients. The problem is that the algorithm is based on a loss that is related to pairwise replacement of two elements, which is not compatible with MQSE. Hence, we introduce an optimization algorithm that combines grid search and iterative methods to estimate the $\widehat{c_{s,t}}$ in MQSE.

To reduce the computational cost of grid search, we assume an approximate product relation $c_{s,t} = \alpha_s \beta_t$. Then, instead of estimating the $\widehat{c_{s,t}}$, we optimize the coupling coefficients α_s of the score function and β_t of the question type asymptotically for $c_{s,t}$ with an iterative method. Specifically, the coupling coefficients $\widehat{\alpha_s}^{(l)}$ and $\widehat{\beta_t}^{(l)}$ are calculated in an alternating iterative way for $l = 1, \cdots$ by the following asymptotic equations.

$$\widehat{\alpha_s}^{(l)} = \underset{\alpha_s}{\operatorname{argmax}} AUPR\left(\sum_{\{s,t\}} \frac{\alpha_s \widehat{\beta_t}^{(l-1)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}, y_{true}\right)$$

$$\widehat{\beta_t}^{(l)} = \underset{\beta_t}{\operatorname{argmax}} AUPR\left(\sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \beta_t}{rank\left(Score_s(\boldsymbol{D}_t)\right)}, y_{true}\right)$$

Finally, the integrated score is calculated from $\widehat{\alpha_s}^{(l)}$ amd $\widehat{\beta_t}^{(l)}$ by the following equation.

$$Score_{MQSE}{}^{(l)}(\boldsymbol{D}) = \sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \widehat{\beta_t}^{(l)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$$

Here, the initial parameter $\widehat{\beta_t}^{(0)}$ is a vector with all elements being 1.

3.4.3 Task-Dependent Rules

In the MQSE learning process and the evaluation process, WISDOM-DX applies task-dependent rules that give priority to the companies with the highest rankings in each industry. Specifically, the task-dependent rules are applied to the following three integration scores.

1) $\sum_{\{s,t\}} \frac{\alpha_s \widehat{\beta_t}^{(l-1)}}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$

2) $\sum_{\{s,t\}} \frac{\widehat{\alpha_s}^{(l)} \beta_t}{rank\left(Score_s(\boldsymbol{D}_t)\right)}$

3) $Score_{ens}{}^{(l)}(\boldsymbol{D})$

Scores (1) and (2) are used to optimize the coupling coefficients, $\widehat{\alpha_s}^{(l)}$ and $\widehat{\beta_t}^{(l)}$, in the asymptotic equations given above. Score (3) is the post-optimization integration score.

In addition, the following rule is helpful for preventing companies in the same industry from dominating the top rankings.

4) Let the score be the reciprocal of the sum of the ranks given by each score and the cost, $\frac{1}{r_{total} + cost\left(r_{seg}\right)}$. We use the following hinge function for $cost\left(r_{seg}\right)$.

$$cost\left(r_{seg}\right) = \begin{cases} a \cdot N\left(r_{seg} - n_{max}\right) & \left(r_{seg} > n_{max}\right) \\ 0 & \left(r_{seg} \leq n_{max}\right) \end{cases}$$

Here, N is the total number of the companies, and n_{max} and a

© 2022 Information Processing Society of Japan

are parameters of the cost function.

4. Experiments

4.1 Purposes

We tested the quality of the company rankings obtained with WISDOM-DX by setting up an evaluation task that was equivalent to a manual survey, along with a baseline method. As mentioned in Section 3.1, the initiatives of excellent DX companies are often reported on the Web. The number of Google searches for DX has been used as an indicator to evaluate trends [61]. Accordingly, we used that number as a baseline method using Web data. Through comparison experiments on the evaluation task, we examined the following points:

1) Feasibility of using Web data

As there is no reported technique for obtaining the appropriate Web information of companies' DX, it is completely unclear how much accuracy can be expected. Hence, by examining the feasibility of using Web data with the baseline method and WISDOM-DX, we obtained initial evidence on this point in relation to using Web data.

2) Validation of QA system and MQSE

We compared the effectiveness of using WISDOM X's answers to 5W1H questions with that of the baseline method. In addition, we evaluated the effectiveness of MQSE, our scoring method based on multiple answers.

3) Validation of WISDOM-DX ranking

We quantitatively evaluated the match between the WISDOM-DX rankings and experts' evaluations regarding the evaluation task. In addition, we examined the validity of the WISDOM-DX ranking by analyzing the top-ranked companies individually with respect to companies that matched the experts' evaluations and those that did not. These results should enable us to identify improvements to WISDOM-DX and provide insights into the functions that companies and research organizations need when using WISDOM-DX.

4) Response from the Evaluated Companies

We surveyed companies on the agreeability and the usefulness of WISDOM-DX. DX promoters in the companies responded to the questionnaire after they were informed of their rankings among the companies of the same industry type as well as the URLs of their answer sets obtained by WISDOM-DX. The results should help us to provide support functions for practical use by companies and research organizations.

4.2 Test Methods

1) Evaluation task

The task was to evaluate 464 companies that responded to the DX Stocks 2021 survey. As described in Section 2.1, the survey resulted in the selection of a total of 48 companies. We refer to these 48 companies as the "DX2021-selected companies," and we evaluated them with the ranking results of WISDOM-DX.

2) Baseline method

The baseline method for automatic ranking with Web data was based on the number of searches in Google Custom Search, a general-purpose search engine. It consisted of an "AND" search of two keywords, "digital transformation" and a company name, together with a ranking of the 464 companies in order of the number of searches.

3) Evaluation measure

We used two evaluation measures: the precision at the break-even point (BEP) and the AUPR. The precision at the BEP was defined as the percentage of the top 48 companies in the ranking that were DX2021-selected companies. The baseline method would be likely to outperform the expected value of 10.3% when 48 companies were randomly selected from 464 companies. Although the precision at the BEP was meaningful in terms of the prediction accuracy for the DX2021-selected companies, it did not evaluate the ranking below 49th place. AUPR, on the other hand, being the area under the precision-recall curve, evaluated the overall quality throughout the rankings. In addition, we investigated achievements other than the DX Stocks 2021 regarding the top 48 companies ranked by WISDOM-DX and the baseline method, i.e., whether they either had won DX-related awards from media or industry organizations or had DX certifications by METI. Regarding the agreeability, usefulness, and necessary functions of WISDOM-DX, we surveyed the evaluated companies with a questionnaire comprising the following questions:

Q1: Does WISDOM-DX provide your company with an agreeable ranking among the companies of the same industry type?

Q2: Can WISDOM-DX be a useful tool for your DX promotion?

Q3: What additional functionality do you need?

4) DX good practices and training data

The DX good practices and the training data were obtained from METI's reports on the companies selected as DX Stocks, Competitive IT Strategy Company Stocks, Noteworthy DX Companies, or Noteworthy IT Strategy Companies from 2015 to 2020 [10]. The good practices consisted of text data introducing DX initiatives of the 255 selected companies. The training data consisted of binary data with values that were positive for the 255 selected companies and negative for other companies.

5) Task-dependent rule

The task-dependent rule here was the hinge-type cost function described as rule (4) in Section 3.4.3. In the experiments, there were $N = 464$ companies. We used parameter values of $n_{max} = 3$ and $a = 0.5$ because they maximized the AUPR for the training data. Note again that this rule was designed to lower the overall rankings of the top two companies within the same industry by adding costs to their scores.

4.3 Test Results

Table 1 lists the results of comparing the top 48 companies ranked by WISDOM-DX and the baseline method. Twenty-seven companies (Group A) ranked by WISDOM-DX were included among the DX2021-selected companies, and 21 companies (Group B) were other than the DX2021-selected companies. In contrast, 11 of the companies (Group A') ranked by the baseline method were included among the DX2021-selected companies, and 37 companies (Group B') were other than the DX2021-selected companies.

Table 2 lists the results of the investigation into the DX

© 2022 Information Processing Society of Japan

Table 1: Results for top 48 companies ranked by WISDOM-DX and the baseline method

	DX2021-Selected Companies	Other Companies
WISDOM-DX	27 (Group A)	21 (Group B)
Baseline Method	11 (Group A')	37 (Group B')

Table 2: Investigation into the Top 48 companies ranked by WISDOM-DX and the baseline method

	DX2021-Selected Companies	Other Companies	
		DX-related awards or DX certifications by METI	None
WISDOM-DX	27 (Group A)	17	4
		21 (Group B)	
Baseline Method	11 (Group A')	20	17
		37 (Group B')	

achievements of other companies. Regarding WISDOM-DX, 17 companies of Group B had won DX-related awards from media or industry organizations or had DX certifications by METI. The remaining 4 companies of Group B did not have any of these awards or DX certifications. Regarding the baseline method, 20 companies of Group B' had won DX-related awards from media or industry organizations or had DX certifications by METI. The remaining 17 companies of Group B' did not have any of these awards or DX certifications. Thus, 91.7% of the top 48 companies ranked by WISDOM-DX and 64.6% of those ranked by the baseline method were taking the initiative in promoting DX at a certain level or higher.

Figure 5 plots the precision-recall curves of the rankings by WISDOM-DX and the baseline method for the evaluation task. Comparison of the two curves shows that WISDOM-DX (denoted as "question-score ensemble" in the figure) outperformed the baseline method ("baseline") at all recall points. The AUPR values were 0.541 for WISDOM-DX and 0.181 for the baseline method. Twenty-seven (Group A) of the top 48 companies obtained by WISDOM-DX matched the experts' evaluation, i.e., the precision (equal to the recall) at the BEP was 56.3%. In contrast, 11 (Group A') of the top 48 companies obtained by the baseline method matched the experts' evaluation, i.e., the precision at the BEP was 22.9%. Thus, WISDOM-DX was superior to the baseline method in terms of both the AUPR and the precision at the BEP.

Table 3 lists the AUPR values for the rankings obtained by each score function of WISDOM-DX and by MQSE. We compared the AUPR values for ranking by calculation of a total of 48 scores for the six question types described in Section 3.2 and the eight score functions described in Section 3.4, and for ranking after integration into a single score by using MQSE. The AUPR with integration by MQSE was 0.541, which was higher than the AUPR for any combination of the six question types and eight score functions.

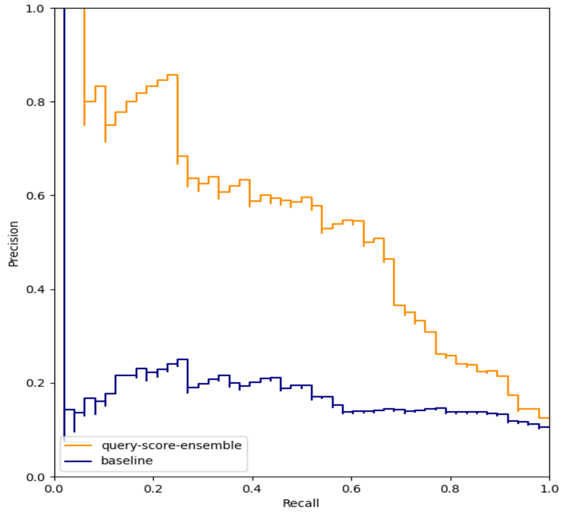


Figure 5: Precision-Recall Curves for WISDOM-DX and the Baseline Method

Table 3: AUPR values obtained by WISDOM-DX

Score Function	Question Type						MQSE
	1	**2**	**3**	**4**	**5**	**6**	
Cnt	0.376	0.414	0.363	0.387	0.378	0.420	**0.541**
Sim	0.396	0.403	0.366	0.397	0.388	0.423	
sim_idf	0.395	0.404	0.369	0.395	0.386	0.425	
sim_tf_idf	0.384	0.402	0.353	0.383	0.380	0.411	
cnt_p	0.400	0.405	0.319	0.433	0.410	0.404	
sim_p	0.401	0.397	0.305	**0.434**	0.398	0.395	
sim_idf_p	0.398	0.404	0.306	**0.434**	0.399	0.398	
sim_tf_idf_p	0.393	0.395	**0.303**	0.428	0.397	0.395	

Table 4: Survey results of agreeability and usefulness

Responses	Q1 (%)	Q2 (%)
Definitely yes	14.3	10.7
Yes, I think so	46.4	35.7
I have no idea	32.1	39.3
No, I don't think so	0	10.7
Definitely no	7.1	3.6

Table 4 lists survey results of 28 companies covering 15 industry types that responded to Q1 and Q2 of the questionnaire. Regarding Q1, 60.7% of the respondents said either "Definitely yes" or "Yes, I think so," 32.1% said "I have no idea," and 7.1% of them said "Definitely no." Regarding the question of Q2, 46.4% of the respondents said either "Definitely yes" or "Yes, I think so," 39.3% said "I have no idea," and 14.3% of respondents

© 2022 Information Processing Society of Japan

said either “No, I don’t think so” or “Definitely no.” Regarding Q3, the provided responses were grouped into the following four types:

1) Diversification of evaluation

・Analysis from multiple perspectives and factor breakdowns would be helpful.

・We would like to see not only the overall ranking, but also the evaluation for each item such as strategy, vision, ability to realize DX, and digital human resources.

・It could be more effective to refine the evaluation period and websites of target companies.

・It would be useful if multiple evaluation axes and item-by-item ratings were provided so that we can link them to our specific actions.

2) More explanation about the results

・It would be helpful to see a list of companies in the same industry type and their rankings for benchmarking.

・It would be useful to see a list of web information and plausibility that WISDOM-DX used for its evaluation.

・Advice on how to improve our ranking would be appreciated.

・It would be helpful to have quantitative measurements as well as the reasons for our ranking.

・It would be even more useful if we could obtain analysis results not only in terms of our own rankings, but also in terms of standard deviations and weaknesses.

3) More companies to be analyzed

・It would improve accuracy if we could specify the URLs to be analyzed.

・It would be very helpful to periodically obtain outgoing DX information from our competitors and stakeholders including foreign information.

4) Other comments

・It is important to enable the PDCA cycle of evaluation in a short cycle for DX promotion.

・The use of in-house analysis tools is not currently being considered.

5. Discussion

5.1 Feasibility of Using Web Data

The precision at the BEP with WISDOM-DX was 56.3%, whereas the precision at the BEP with the baseline method was 22.9% and the expected value was 10.3% with random selection. As mentioned in Section 4.3, 91.7% of the top 48 companies ranked by WISDOM-DX were taking the initiative in promoting DX at a certain level or higher, compared to 64.6% of those ranked by the baseline method. Although 4 companies of the top 48 companies ranked by WISDOM-DX had no significant achievements related to DX on their own, they were working with other companies on DX and supporting the DX initiatives of other companies. We will need to examine how to evaluate initiatives with other companies, and it should be possible to improve the accuracy of WISDOM-DX by identifying both a company’s own activities and its joint activities with other companies from Web data. These results show that WISDOM-DX offers more promising performance than the baseline method, and demonstrate the feasibility of automatically evaluating DX initiatives using Web data.

5.2 Validation of QA System and MQSE

It was confirmed that the Web information obtained by WISDOM-DX with 5W1H questions covered the content related to the six questionnaire items mentioned in Section 2.1, as follows.

・Question Types 1 (how) and 5 (why):
Management vision and business model; strategy; use of digital technology and IT systems to implement strategy.

・Question Type 2 (where):
Organization and scheme to implement strategy.

・Question Type 3 (who):
Organization and scheme to implement strategy; governance.

・Question Type 4 (what):
Sharing of results and key performance indicators.

・Question Type 5 (when):
Not directly for specific items, but for general purposes.

5W1H modeling was helpful for obtaining DX-related information from Web data.

The AUPR obtained from the MQSE score of WISDOM-DX was 0.541, which was about three times higher than the AUPR of 0.181 for the baseline method. Even without MQSE, the AUPR values for WISDOM-DX were still higher than those of the baseline method: they ranged from 0.303 to 0.434, as seen in Table 3. In other words, WISDOM-DX, which uses 5W1H questions and WISDOM X answers on companies' DX initiatives, achieved a higher accuracy in ranking than the baseline method, which used Google Custom Search, a general-purpose search engine. Hence, these results demonstrate that WISDOM X is a QA system that can provide more helpful DX-related information on companies than Google Custom Search can. In addition, the AUPR of 0.541 with score integration, which was higher than the highest value of 0.434 without MQSE, confirmed that MQSE is effective in improving the ranking accuracy.

5.3 Validation of WISDOM-DX Ranking

Table 5 lists the DX selection types and industry types of Group A, comprising 27 DX2021-selected companies that were included among the top 48 companies ranked by WISDOM-DX. For the DX selection types, the DX Stocks 2021 are indicated by ●, and the DX Noteworthy Companies 2021 are indicated by ▲. Among the 27 companies, the average rank of the DX Stocks 2021 was 11.5, and that of the Noteworthy DX Companies 2021 was 16.3. Thus, in the WISDOM-DX rankings, the DX Stocks 2021 were evaluated more highly on average than the Noteworthy DX Companies 2021 were. That is, the DX Stocks 2021 were distributed near the top, while the DX Noteworthy Companies 2021 were distributed near the bottom. The industry types of the top five companies were Electric Appliances and Information & Communication. The Grand Prix companies were included in the top five. Even though we used the training data to optimize the coupling coefficients as positive and negative values without distinguishing the DX Stocks and DX Noteworthy companies, the selection types of the companies that matched the WISDOM-DX results were almost entirely consistent with the evaluation results

© 2022 Information Processing Society of Japan

Table 5: Selection types and industry types for Group A

	Stocks	Noteworthy	Industry Type
1	●		Electric Appliances
2		▲	Information & Communication
3		▲	Electric Appliances
4	●		Information & Communication
5	●		Electric Appliances
6	●		Other Products
7	●		Air Transportation
8	●		Land Transportation
9		▲	Wholesale Trade
10	●		Machinery
11		▲	Wholesale Trade
12	●		Chemicals
13		▲	Services
14		▲	Real Estate
15	●		Rubber Products
16		▲	Insurance
17		▲	Air Transportation
18	●		Pharmaceuticals
19		▲	Retail Trade
20	●		Retail Trade
21	●		Land Transportation
22		▲	Machinery
23	●		Wholesale Trade
24		▲	Foods
25		▲	Glass and Ceramics Products
26		▲	Banks
27		▲	Pharmaceuticals

for the DX Stocks 2021 and the DX Noteworthy Companies 2021. As described in Section 2.1, one or two companies were selected as DX Stocks 2021 for each of the 33 industry types [25]. Even if there were three or more highly evaluated companies in the same industry, it was not possible to select all of them as DX Stocks 2021. In that case, the remaining companies could be selected as DX Noteworthy Companies 2021. Accordingly, we cannot rule out this possibility for the second ranked Information & Communication company and the third ranked Electric Appliances company.

For the 21 DX2021-selected companies (Group C) that were not among the top 48 companies ranked by WISDOM-DX, Table 6 lists the causes of mismatch, as described below, the number of companies for each cause, and our plans to improve the performance in these cases.

1) Insufficient answers

Thirteen of Group C received a total of less than 10 answers from WISDOM X for all question types, and eight of the 13 companies received no answers. The number of searches for these 13 companies with the baseline method ranged from 4,270 to 4,501,000, which suggests that there is relevant information in Web data. The main causes that prevented WISDOM X from obtaining enough answers were insufficient crawling of Web data and an inability to extract relevant Web data because of insufficient questions. The former problem can be solved by having WISDOM X crawl Web pages that are retrieved by the baseline method. The latter problem can be mitigated by expanding the questions with additional entries for aliases and synonyms in the domain dictionary: in these experiments, only a few company abbreviations were registered there. By expanding the questions, we should be able to increase the number of answers.

2) Disparity among industry types

Four of Group C were ranked first or second within their industries. As described in Section 3.4.3, the task-dependent rules give priority to the companies with the highest ranking in each industry. However, the first-ranked company in a certain industry was sometimes ranked lower than the second- and third-ranked companies in other industries. Accordingly, we could improve the performance by increasing the priority of the first position within each industry type.

3) Difference in evaluation period

Four companies were probably in Group C because of differences in the evaluation period. Specifically, newly established companies were ranked lower because they had less Web data available than older companies. Companies with less Web data are more likely to get lower scores related to answer volumes. In contrast, some companies with large answer volumes because of their accumulated past achievements were ranked higher than companies with superior recent DX initiatives. If the evaluation includes all past achievements, the same companies will be ranked highly every year. We can overcome this issue by normalizing the aggregation period for Web data and considering the freshness of the answer volumes.

5.4 Response from the Evaluated Companies

1) Agreeability

Regarding agreeability, 60.7% of respondents offered positive comments because they understood the situations of the same industry type through a kind of benchmarking, while 32.1% of them had no idea because they had little information about the

Table 6: Mismatch causes and improvement plans for Group C

Cause of Mismatch	Improvement Plan	Number of Companies
Insufficient answers by WISDOM X	・Crawl more Web pages with WISDOM X ・Expand questions with additional entries of aliases and synonyms in domain dictionary	13
Disparity among 33 industry types	・Increase priority of first position within industry type	4
Difference in evaluation period	・Account for period of Web data and freshness of answer volumes	4

© 2022 Information Processing Society of Japan

relative status of their own companies. As for the others, 7.1% offered negative comments because the ranking algorithm was not clear to them even though Web pages obtained by WISDOM-DX were presented as the evidence data used for the algorithm. It is necessary to improve how the evidence for rankings is presented.

2) Usefulness

Regarding usefulness, 46.4% of the respondents offered positive comments because they could identify time-series changes of their relative position from the objective viewpoints, while 39.3% of them had no idea because they could not understand how their rankings were calculated. The others, 14.3%, offered negative comments because they could not see their improvement points through their rankings and the Web pages obtained by WISDOM-DX. It is necessary to evaluate their DX situations from the viewpoints of their interests.

3) Necessary Functions

Regarding the necessary functions, the provided responses were related with diversification of evaluation, more explanation about the results, and more companies to be analyzed. We will implement the support functions as described in Section 6.

6. Future Issues

We plan to improve WISDOM-DX as described above so that it can support DX evaluations by companies and research organizations. The following functions will be provided as support tools for DX promoters and evaluators.

1) Diversification of evaluation through question expansion

WISDOM-DX will provide an interface that allows users to edit evaluation items. When a company promotes DX as a management strategy, it should adapt this effort to match its own situation in analyzing its own issues and those of its competitors. The interface will enable users to analyze their particular interests by easily registering keywords in the domain dictionary. For example, countermeasures against COVID-19, AI utilization, cloud utilization, third-party collaboration, and human resource development are topics of high interest for companies promoting DX. When these topics are registered in the domain dictionary, WISDOM-DX generates the following questions for Company A:

・What did Company A implement for countermeasures against COVID-19?

・What did Company A implement for AI utilization?

・What did Company A implement for cloud utilization?

・What did Company A implement for third-party collaboration?

・ What did Company A implement for human resource development?

WISDOM-DX performs the same analysis for other companies, including Company A's competitors. The visualized results, as shown in Figure 6, are useful for multifaceted analysis and evaluation by Company A.

2) Integrated use with tools for summarization and text mining

WISDOM-DX will also provide users with an interface that integrates answer sets with tools for summarization and text mining. For example, when Company A obtains answer sets for its competitor on countermeasures against COVID-19, AI utilization, cloud utilization, third-party collaboration, and human resource development, the tools will make it easy to understand the contents of the competitor's initiatives. The good practices of other companies have been reported to be helpful for DX promotion activities such as business strategy planning [30]. Integrated use with these tools for summarization and text mining will enhance WISDOM-DX's capability because it will make companies' DX efforts more visible.

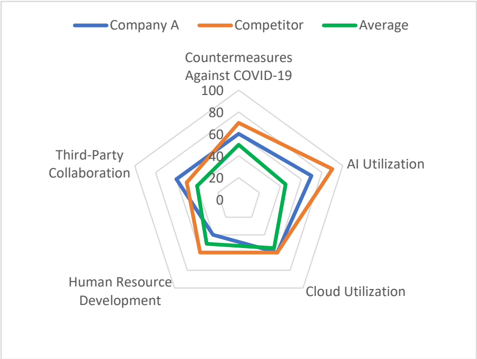


Figure 6: Analysis Example with a Radar Chart

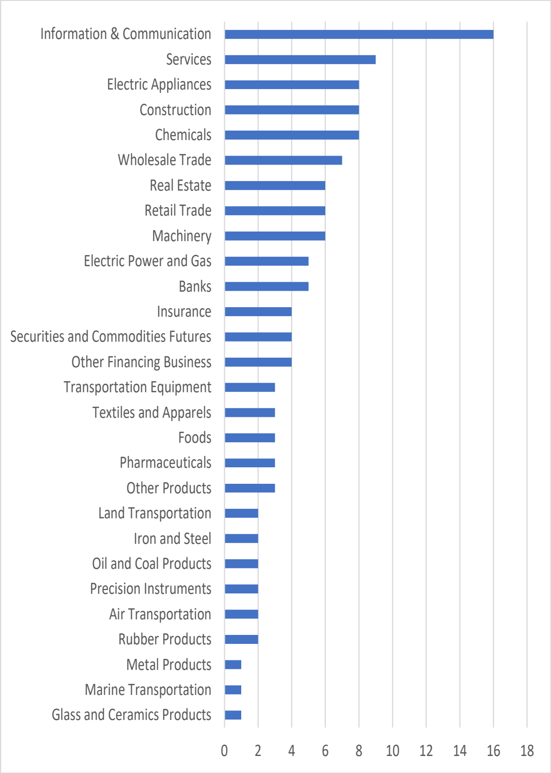


Figure 7: Number of DX Stock Companies by Industry from 2015 to 2021

These tools can also be used by research organizations to identify characteristics and trends by industry types. In the surveys for DX Stocks and Competitive IT Strategy Company Stocks that were conducted from 2015 to 2021, a total of 303 companies (126 different companies) were selected as DX Stocks or Noteworthy Companies. Figure 7 shows the number of companies by industry type. Our tools can support the evaluation and planning of DX promotion programs by presenting a summary of the characteristics of each industry in which many companies are selected or few companies are selected.

© 2022 Information Processing Society of Japan

3) Expansion of Web Data

The capability to include and evaluate multilingual Web data will make it possible to evaluate the DX initiatives of non-Japanese companies. As reported in the DX White Paper 2021 [12], Japanese companies have lagged behind U.S. companies in their DX initiatives. Specifically, about 56% of companies in Japan are involved in DX, compared to about 79% of companies in the U.S. Likewise, the percentage of companies that are not engaged in DX is 33.9% in Japan but only 14.1% in the U.S. Hence, WISDOM-DX will facilitate scoring and cross-analysis through English and multilingual Web data. It will also support research aimed at improving international competitiveness by evaluating both Japanese companies and overseas companies.

7. Conclusion

To make it possible for companies to make objective evaluations, we developed WISDOM-DX, a system that leverages a QA system based on Web information that automatically evaluates the DX initiatives of companies without manual DX survey. By modeling designers' evaluation items in the form of 5W1H (when, who, where, what, why, how) questions, WISDOM-DX evaluates DX initiatives by scoring an answer set generated by the QA system.

To examine the feasibility of using Web data, WISDOM-DX and a baseline method that used Google Custom Search were evaluated by ranking 464 companies that responded to the DX Stocks 2021 survey from which DX experts selected 48 companies for distinction as DX Stocks 2021 or Noteworthy DX Companies 2021. Regarding the top 48 companies ranked by WISDOM-DX, 27 of them were included among the 48 selected companies and 17 of them had received DX-related awards or certifications, indicating 91.7% had a certain level of achievement for their DX initiatives. In contrast, 11 of the top 48 companies ranked by the baseline method were included among the 48 selected companies and 38 of them had received DX-related awards or certifications, indicating 64.6% had a certain level of achievement for their DX initiatives. When WISDOM-DX and the baseline method were evaluated for searching for the 48 selected companies, the area under the precision-recall curve (AUPR) values obtained by WISDOM-DX and the baseline method were 0.541 and 0.181, respectively. In addition, the respective precision values were 56.3% and 22.9%. The survey of WISDOM-DX with the questionnaire to the evaluated companies showed that 60.7% offered positive responses and 32.1% neutral responses regarding the agreeability of their rankings, and that 46.4% offered positive responses and 39.3% neutral responses regarding the usefulness of the system. These results show that WISDOM-DX had more promising performance than the baseline method, and that it offers the prospect of automating large-scale analysis and evaluation of DX initiatives as a first step in using Web data for benchmarking companies.

Our experiments demonstrated the need to improve the QA system's accuracy by identifying both a company's own activities and its activities in collaboration with other companies, by increasing the number of answer sets, and by accounting for differences among industry types and evaluation periods. We will implement these improvements so that WISDOM-DX will be useful for a wider variety of companies and research organizations.

Acknowledgments

We express our thanks to all the personnel related to WISDOM-DX. We are especially indebted to Mr. Hirohide Hirai, Mr. Jun-ichi Eguchi, Mr. Takefumi Tanabe, and Mr. Takuya Watanabe of METI for their encouragement and cooperation, and to Dr. Kiyotaka Uchimoto and Dr. Masahiro Tanaka of NICT for their tremendous support. We are also grateful to the late Prof. Mikio Aoyama of Nanzan University for his invaluable contributions to DX in Japan, and we offer our deepest condolences for his loss.

References

[1] Okumura, A.: Editor's Message to Special Issue on Digital Architecture Design, Information Processing, Vol.62, No.6, pp. 284-287. (May 15, 2021) (in Japanese)

[2] Digital Agency: Priority plans for the formation of a digital society, (Dec.24,2021) (in Japanese) https://cio.go.jp/sites/default/files/uploads/documents/digital/20211224_policies_priority_doc_01.pdf

[3] Kobayashi, Y.: Current Status and Issues of Evidence-Based Policy Making in Japan, Japanese Journal of Evaluation Studies, Vol. 20, No.2, pp.33-48. (Jul. 2020) (in Japanese)

[4] Ministry of Economy, Trade and Industry: Guidelines for Promotion of Digital Transformations Formulated. (Dec. 12, 2018) https://www.meti.go.jp/english/press/2018/1212_003.html

[5] Ministry of Economy, Trade and Industry: DX Report - Overcoming the IT System "Cliff of 2025" and Full-Scale Development of DX, (Sep.7. 2018) (in Japanese) https://www.meti.go.jp/press/2018/09/20180907010/20180907010-3.pdf

[6] Ministry of Economy, Trade and Industry: DX Report 2 (interim report), (Dec.24. 2020) (in Japanese) https://www.meti.go.jp/press/2020/12/20201228004/20201228004-2.pdf

[7] Ministry of Economy, Trade and Industry: DX Report 2.1 （DX Report 2 supplementary edition）,(Aug.31. 2021) (in Japanese) https://www.meti.go.jp/press/2021/08/20210831005/20210831005-2.pdf

[8] Ministry of Economy, Trade and Industry: Digital Governance Code，(Nov.9. 2020) (in Japanese) https://www.meti.go.jp/shingikai/mono_info_service/dgs5/pdf/20201109_01.pdf

[9] Ministry of Economy, Trade and Industry: DX Certification System (certification system based on Article 31 of the Act on the Promotion of Information Processing)，(Nov.9.2020) (in Japanese) https://www.meti.go.jp/policy/it_policy/investment/dx-nintei/dx-nintei.html

[10] Ministry of Economy, Trade and Industry: Digital

© 2022 Information Processing Society of Japan

Transformation Stocks (DX Stocks) 2021. (Jun.7, 2021) (in Japanese)
https://www.meti.go.jp/policy/it_policy/investment/keiei_meigara/dx-report2021.pdf

[11] IPA Information-technology Promotion Agency, Japan : The DX White Paper 2021, US-Japan Comparative Study on DX Strategy, Human Resources, and Technology, (Dec.1. 2021) (in Japanese)
https://www.ipa.go.jp/ikc/publish/dx_hakusho.html

[12] Ministry of Economy, Trade and Industry: "DX promotion index" and its Guidance. (Jul., 2019) (in Japanese), https://www.meti.go.jp/press/2019/07/20190731003/20190731003-1.pdf

[13] Ministry of Economy, Trade and Industry: Study Group for Acceleration of Digital Transformation WG1 Plenary Report, (Dec.28. 2020) (in Japanese)
https://www.meti.go.jp/press/2020/12/20201228004/20201228004-4.pdf

[14] Sato, I.: The Art and Science of Qualitative Data Analysis: What We Can and Cannot Do with QDA Software, The Japan Institute of Labour, Vol.57, No.12, pp81-96. (2015-12) (in Japanese)

[15] Machida, K.: Advantages and points to consider when applying quantitative text mining to qualitative Japanese text data, SCU Journal of Design & Nursing Vol. 13, No. 1, pp.47-53. (2019) (in Japanese)

[16] Okabe, D.: Data Analysis for Communication Studies, Tazaki, K., ed., Data analysis for communication studies. Nakanishiya Shuppan, pp.189-201. (Sep. 2015) (in Japanese)

[17] Nikkei BP Intelligence Group: Digital Transformation II. (Nov. 25, 2020) (in Japanese)
https://info.nikkeibp.co.jp/nxt/campaign/b/279660/

[18] Yano Research Institute Ltd.: Digital Transformation (DX) Market 2020. (Jul. 2020) (in Japanese)
https://www.yano.co.jp/press-release/show/press_id/2487

[19] International Data Corporation Japan: Released the results of a survey on digital transformation trends among Japanese companies. (Dec. 7, 2020) (in Japanese)
https://www.idc.com/getdoc.jsp?containerId=prJPJ47071820

[20] Japan Users Association of Information Systems: Enterprise IT Trends Survey 2021. (Apr. 28, 2021) (in Japanese)
https://juas.or.jp/library/research_rpt/it_trend/

[21] Japan Information Technology Service Industry Association: Contribution of Information Service Companies to the Digital Transformation (DX) of Society. (May 30, 2019) (in Japanese)
https://www.jisa.or.jp/publication/tabid/272/pdid/30-J007/Default.aspx

[22] Japan CTO Association: DX Trend Survey Report 2021 Edition. (Apr. 12, 2021) (in Japanese) https://cto-a.org/news/2021/04/12/4956/

[23] Tokyo Metropolitan Government: DX Promotion Pilot Project (Phase 1). (Mar. 19, 2021) (in Japanese)
https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2021/03/19/05.html

[24] Kanagawa Prefectural Government: Kanagawa DX Project Promotion Project. (May 17, 2021) (in Japanese)
https://www.pref.kanagawa.jp/docs/sr4/dx-project.html

[25] Ministry of Economy, Trade and Industry: "Video of the announcement of companies selected for Digital Transformation Stocks (DX Stocks) 2021 is now available!" (Jul.13. 2021) (in Japanese)
https://www.meti.go.jp/press/2021/07/20210713005/20210713005.html

[26] Ministry of Economy, Trade and Industry: DX Stocks/Competitive IT Strategy Company Stocks (Jun. 11, 2021) (in Japanese)
https://www.meti.go.jp/policy/it_policy/investment/keiei_meigara/keiei_meigara.html

[27] Ministry of Internal Affairs and Communications: Information and Communications in Japan 2021 White Paper: Present Status and Challenges for Digital Transformation in Corporate Activities
https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2021/chapter-1.pdf#page=7

[28] Ogi, S.: Technical Report about Text Mining, Journal of the Japanese Society of Computational Statistics, Vol.28 No.1, pp.31-40. (2018) (in Japanese)

[29] Ishida, M. and Kim, M., eds.: Corpus and Text Mining, Kyoritsu Shuppan Co., Ltd., pp.1-14. (2012) (in Japanese)

[30] Goto, M., Harada, S., Tanabe, W.: Construction of Analogy Evaluation Model based on Structured Strategy Map, Proceedings of the 2007 Fall Conference of Japan Society for Management Information. (2007) (in Japanese)

[31] Nikkei Business Publications, Inc.: Nikkei Business, Dec.11, 2000 – Sep.1, 2003. (in Japanese)

[32] Zixuan Ke, Vincent Ng: Automated Essay Scoring: A Survey of the State of the Art, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Survey track. Pages 6300-6308. https://doi.org/10.24963/ijcai.2019/879

[33] Attali, Y. and Burstein, J.: Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment. (2006)

[34] Ishioka, T., Kameda, M.: JESS : An Automated Japanese Essay Scoring System, Journal of the Japanese Society of Computational Statistics, Vol.16, No. 1, pp. 3-19. (Dec. 2003) (in Japanese)

[35] Nakajima, K.: Automated Classification of Short Answers using Supervised Machine Learning, In Proceedings of the 26th Annual Conference of Japan Society for Educational Technology, pp.639-640. (2010) (in Japanese)

[36] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631-1642. (2013)

[37] Kim, Y.: Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751. (2014).

[38] Terada, R., Kubo, A., Shibata, T., Kurohashi, Y. and Okubo, T.: Automated Short-Answer Grading using Neural Networks, In Proceedings of the 22nd Annual Conference of Natural Language Processing. pp.370-373. (2016) (in Japanese)

[39] Strzalkowski, T., Harabagiu, S.: Advances in Open Domain Question Answering. Springer. (Oct. 2006)

[40] The New York Times: Computer Wins on 'Jeopardy!': Trivial, It's Not, (Feb. 16, 2011)
https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

[41] Sewon Min, et all: NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned, (Jan. 2021)
https://colinraffel.com/publications/arxiv2021neurips.pdf

[42] Mihara, E., Fujii, A., Ishikawa, T.: A Helpdesk-oriented Question Answering System Using the World Wide Web, Proceedings of the 5th Forum on Information Technology 2005, pp.163-166. (Aug. 2005) (in Japanese)

[43] Aizawa, Y., Tsuchiya, Y., Watabe, H.: Consideration of question answering system based on Wikipedia, Proceedings of the 18th Forum on Information Technology, pp.165-166. (Aug. 20, 2019) (in Japanese)

[44] National Institute of Information and Communications Technology

© 2022 Information Processing Society of Japan

（NICT）: What is WISDOM X ? (in Japanese) https://www.wisdom-nict.jp/#top

[45] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer.: A semi-supervised learning approach to why-question answering. In Proceedings of AAAI-16, pp. 3022–3029, (2016).

[46] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, Istv´an Varga, Jong-Hoon Oh, and Yutaka Kidawara: Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In Proceedings of ACL 2014, pp. 987–997, (2014).

[47] John-Hoon Oh, Kazuma Kadowaki, Julien Kloetzer, Ryu Iida and Kentaro Torisawa: Open domain why-question answering with adversarial learning to encode answer texts. In Proceedings of ACL 2019, pp. 4227– 4237, (2019).

[48] Kazuma Kadowaki, Ryu Iida Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer: Event causality recognition exploiting multiple annotators' judgments and background knowledge. In Proceedings of EMNLP 2019, pp. 5820-5826, (2008).

[49] John-Hoon Oh, Ryu Iida, Julien Kloetzer, and Kentaro Torisawa: BERTAC: Enhancing transformer-based language models with adversarially pretrained convolutional neural networks. In Proceedings of ACL-IJCNLP 2021, pp. 2103– 2115, (2021).

[50] MasahiroTanaka, KenjiroTaura, and KentaroTorisawa.: Autonomic resource management for program orchestration in large-scale data analysis. In Proceedings of IPDPS 2017, pp.1088-1097. (Jun.30 2017).

[51] Impress: Impress DX Awards, Who are the leaders in digital transformation? (Oct.2020) (in Japanese) https://dx-awards.impress.co.jp/

[52] NIKKEI Computer：IT Japan Award 2021，(Jun. 8. 2021) (in Japanese) https://www.nikkeibp.co.jp/atcl/newsrelease/corp/20210608/

[53]Japan Institute of Information Technology: Selected as the winner of the 39th IT Award for 2021, (Feb. 2022) (in Japanese) https://www.jiit.or.jp/im/award.html

[54] NIKKEI X TREND: The "Meaning of DX" is classified by type of industry, (Dec.1. 2020) (in Japanese) https://xtrend.nikkei.com/atcl/contents/18/00382/00005/

[55] Honnibal M, Montani I: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, (2017).

[56] Asahara M., Kanayama H., Miyao Y., TanakaT., Omura M., Murawaki Y., Matsumoto Y.: Japanese Universal Dependencies Corpora, Journal of Natural Language Processing, vol. 26, no. 1, pp.3-36, (2019). (in Japanese)

[57] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods", SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 758-759, (Jul. 2009).

[58] Voorhees, E. M., and Harman, D. K., Eds. TREC – Experiment and Evaluation in IR. MIT Press (2005).

[59] Zhou, Z.: Ensemble Methods: Foundations and Algorithms, Machine Learning & Pattern Recognition Series, CRC Press, (2012).

[60] Joachims, T.: A Support Vector Method for Multivariate Performance Measures, Proceedings of the 22nd International Conference on Machine Learning, ACM Press, pp. 377-384, (2005).

[61] Ebata, H.: DX from Marketing Perspective, NIKKEI BP, (Oct. 2020).(in Japanese)

© 2022 Information Processing Society of Japan