Use of AI in examining annual reports

Summary of the proof-of-concept test results

Financial Services Agencies "Open Policy Lab"

September 2019

The project "Use of AI in examining annual reports"

- The FSA started **"Open Policy Lab"** as a framework that encourages young staff members to proactively make new policy proposals. It aims to enhance human resource development and is expected to facilitate unique ideas regarding policies.

- As one of the projects of this "Open Policy Lab," volunteers launched a project entitled **"Use of AI in examining annual reports."**

- As **public interest in narrative information in annual reports is growing**, the project examined if technologies including AI could be used to **effectively and efficiently examine annual reports** and be **good ways to find better narrative information**.

Recruit cooperation

■ What can be done about the analysis of Japanese texts at present?

The project recruited companies to cooperate with a PoC (Proof-of-Concept) test using AI on the FSA's website (May 8, 2019).

(Outline of the test)

- ➢ The project examined if AI, instead of a human, could judge if descriptions in annual reports were good or bad.
- ➢ The project tested;
  - whether AI could read descriptions of specific items in annual reports, and
  - whether AI could find similar features in other annual reports
- ➢ This test ended on June 30, 2019 (Test period is about **one month**)

➡ **20 companies (18 groups) cooperated in the project.**

- ✓ Information & Communication
- ✓ Think tanks
- ✓ AI ventures/AI start-ups
- ✓ Disclosure/IR
- ✓ Accounting firms

List of companies participating in the project

1. Information & Communication (eight companies)

NTT DATA, NT DATA CCS, QUICK/Hitachi (joint), Systems Engineering Consultants, IBM Japan, Japan third party, Microsoft Japan/ Persol Process ＆Technology (joint), Fujitsu

2. Think tank (one company)

Daiwa Institute of Research

3. AI ventures/ AI startups (six companies)

Arithmer, ZAISAN Net, JIAI, Deep Data Research, B2B Makers, MILIZE

4. Disclosure/IR (one company)

TAKARA PRINTING

5. Accounting firms (two firms)

KPMG AZSA, EY Advisory＆Consulting

Summary of PoC test (Illustrative)

Annual Reports (Description regarding "management indicators")

All data for the past five years can be freely acquired via EDINET (API).
***See next slide for details**

Judged and Classified
by Humans

Best practice

Examples that serve as a wide reference with extensive descriptions.

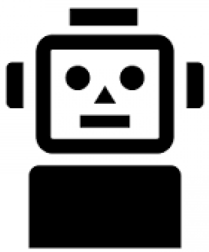
Need to improve

Insufficient description of management indicators, management analysis in light of the indicators, etc.

Others

Normal cases.

The project was aiming to determine if



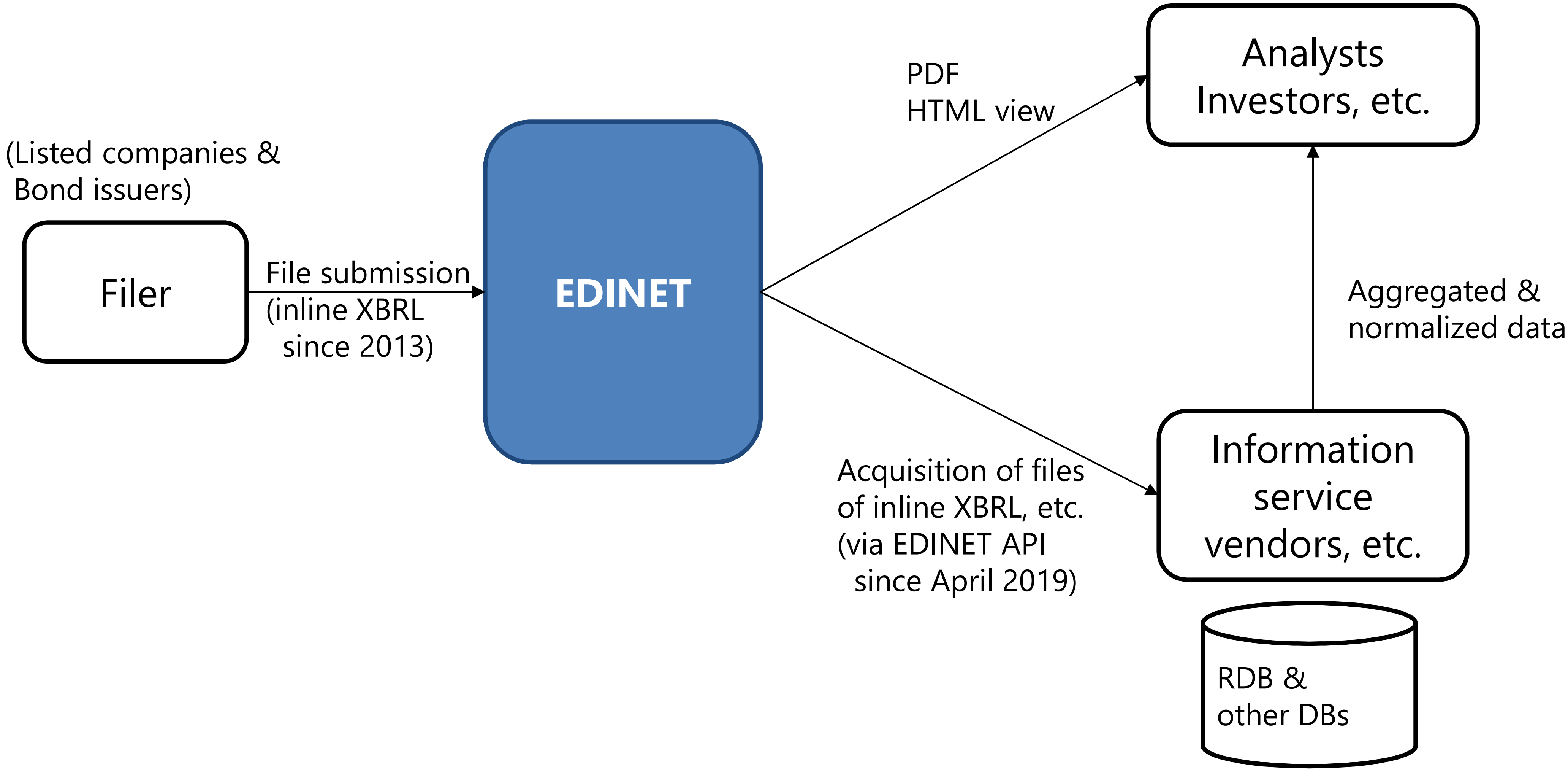
it is possible for AI to make the same judgments and classifications as humans
by making the AI learn about annual reports and the data of the above examples classified by the FSA.

(Appendix) Overview of EDINET

Electronic **D**isclosure for **I**nvestors' **NET**work

- **Electronic disclosure system under the Financial Instruments and Exchange Act**
  - Mandatory use for filing annual report, quarterly report, internal control report, securities registration statement, shelf registration statement, tender offer notification, large volume holding report, etc.
- **Accessible through the internet to file or browse the disclosure documents**
- **Number of filers (annual report):**
  - **Approx. 4,300 companies (mainly listed companies)**
  - **Approx. 4,700 investment funds**

(Appendix) Data flow of EDINET



Main approaches of the test

The following are the four main approaches conducted by companies:

JFSA provided data set

Rule-based

A human programs the classification rules

Machine learning

The machine learns classification rules from data

Deep learning

The machine extracts features from data automatically

Overall analysis

The machine analyzes features / trends

Classification / Analysis results

Evaluate

- Degree of conformity with manual classification
- Has it been classified without omission?
- Features and trends

Main approaches of the test（Rule-based）

JFSA provided data set

Rule-based

- Create a classification flow by programming based on the rules (such as the Cabinet Office Ordinance) and examples.
- Prepare a dictionary by extracting KPI expressions.
- Learn the concept of KPI phrases and learn rules.
- Understand the Japanese syntactic structure, the chapter structure and table structure of the document.
- Extract information by natural language processing and perform statistical analysis.
- Check the existence of charts and the similarity of character strings by using XBRL data.

Classification / Analysis results

Evaluate

Main approaches of the test（Machine learning）

JFSA provided data set

Machine learning

1. Supervised learning

Training data → Classification model

Build a classification model using teacher data. The verification data was classified by using the classification model.

- ✓ Text data was converted into numerical expressions, and a classification model was constructed by applying an algorithm.
- ✓ Quantification method: Vectorization from the number of appearances of words (BOW) / Expression of importance from the appearance frequency of words in documents (TF-IDF) / Topic extraction from word distribution (Topic model) / Vectorization of word meaning (Word2Vec), etc.
- ✓ Algorithm: Random forest, logistic regression, etc.

2. Unsupervised learning

Learning from data and grouping (such as clustering) without giving teacher data corresponding to correct answers

Classification / Analysis results

Evaluate

Main approaches of the test（Deep learning）

JFSA provided data set

Deep learning

Automatic feature extraction from data.
Deep learning is a type of machine learning. The machine itself extracts the feature even if the person does not specify the point (feature) to be noted in the data.

- As preprocessing, the sentence was divided into words and specific parts of speech are removed.
- Convert text data into a numerical representation and apply the algorithm (mathematical calculation procedure) to construct a classification model
  - ✓ Digitization method: Vectorization of word meaning (Word2Vec, Japanese BERT model), textualization (Doc2Vec)
  - ✓ Algorithm: Neural network (three layers, five layers), bidirectional LSTM, CNN

Classification / Analysis results

Evaluate

Main approaches of the test（Overall analysis）

JFSA provided data set

Overall analysis

Analyzes the features and trends of descriptions that appear in best practice and the other examples.

- Comparison of character count, word count, and word appearance rate
- Classification by extracting feature words of the sentence from the appearance frequency of words (TF-IDF)
- Negative-positive analysis using a polarity dictionary (a dictionary that evaluates the positive / negative degree of a word)
- Extract word features of sentences (Word2Vec) and investigate similarity
- Correlation with head office location

Classification / Analysis results

Evaluate

Test results and issues (1/5)

Models and Approaches

- In rule basis, classification rules can be made based on clear logic, and it can achieve high accuracy by implementing human determination logic. But tuning and maintenance by a human is needed.
- In machine learning and deep learning, even if AI can classify data the same as teacher data, it is not easy to clarify why it is appropriate and where it is not appropriate.
- In machine learning, when the learning object is a long sentence, noise tends to increase and the learning effect tends to decrease. In order to grasp the features of the description by machine learning, it is necessary to specify the description part in detail.
- In the description information of annual reports, the same noun expression is seen in many documents, and there is not much difference, so it was difficult to grasp the characteristics with probability theory and statistical analysis methods. There is a possibility that effective analysis can be achieved by converting the document structure into features by natural language processing and capturing the features after classification according to the type of industry.

Test results and issues (2/5)

Deep learning

- Deep learning can automatically classify the data without human support, but it did not lead to a good result in the project. It is considered possible that deep learning requires more teacher data than machine learning, and is not suitable for identifying a document that lacks the necessary descriptions.

- However, there is a possibility that accuracy can be improved by using a method that compensates for the shortage of teacher data (for example, learning and adapting specific words and contexts from a large amount of past annual reports [transfer learning] and other methods).

Test results and issues (3/5)

Training data related to machine learning and deep learning

- In order to improve accuracy, it is necessary to accumulate more teacher data with the accurate answer label.
- It is important to maintain a dictionary (corresponding to management, accounting, auditing, laws, abbreviations, etc.) related to management indicators and technical terms described in annual reports.
- Even if the dictionary is elaborated upon, AI may not be able to handle unknown management indicators (for example, qualitative indicators such as productivity improvement and supply chain strengthening).
- If the teacher data is clear, accuracy can be expected to improve, but if there is a fluctuation of human judgment such as a qualitative description, machine learning may not work.

Test results and issues (4/5)

Best practice

- Many AIs could not distinguish the “best practice” very well due to a shortage of cases. But in some cases, AI learned about a specific disclosure pattern (such as how management has set management indicators) and extracted information by natural language processing and performed a statistical analysis that led to more efficient “best practice” predictions.

- There were also cases where AI identified the word that is the topic of the sentence by analyzing the number of occurrences of the word. “Best practice” cases use many proper and concrete words, which explain their services.

Test results and issues (5/5)

Others

- In the descriptive information of annual reports, technical terms are frequently used and adjectives are rarely used. Focusing on nouns would enable better understanding in order to analyze features.

- When frequent words were extracted from the disclosed examples, words used for specific analysis such as “increase / decrease,” “change,” and “loss” were found in “best practice.” On the other hand, in another group of cases, ambiguous expressions such as “try” and “plan” were seen.

- Regarding use of XBRL data, we saw results with extracting fixed descriptions by calculating the similarity rate with the previous year for descriptions of management policies, etc.

- Statistical information of annual reports (number of characters, number of words, frequency of use for each word, negative/positive analysis using a polar dictionary, head office location, etc.) functions as a reference when evaluating companies from multiple perspectives.

Summary of the results

Our Insights

Progress of NLP technology

- Though narrative information is much more difficult to read and analyze by computer than financial information in general, the companies participating in the project achieved reasonably good results using various approaches.
- This result suggests that **text information could be widely analyzed precisely and in large quantities by AI** thanks to the progress of natural language processing (NLP) technology.

Possibility of AI

- It is seen in some approaches that AI understands the sentence structure and context, and try to analyze sentences in **a more human-like way**.
- If there was **sufficient time** and **more data to be analyzed**, AI would be able to **work much better**. On the other hand, since the criterion of AI tends to be a black box, it is important to consider how to use it.

Role of humans and AI

- Many companies emphasized that it was vital to develop AI **jointly with the person who judged the evaluations of descriptions**.
- Now AI is not all-in-one. As humans and AI have their own areas of expertise, **it is important for humans and AI to play their own roles** to achieve better performance. Humans should provide their judgment and knowledge to AI as well as interpretation of the results and feedback.