Automated Generation of Coding Rules: Text-Mining Approach to ISO 26000

Tetsuya Nakatoh*, Satoru Uchida†, Emi Ishita‡ and Toru Oga§
*Research Institute for Information Technology, Kyushu University.
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, JAPAN
Email: nakatoh@cc.kyushu-u.ac.jp
†Faculty of Languages and Cultures, Kyushu University.
Email: uchida@flc.kyushu-u.ac.jp
‡Research and Development Division, Kyushu University Library.
Email: ishita.emi.982@m.kyushu-u.ac.jp
§Faculty of Law, Kyushu University.
Email: toga@law.kyushu-u.ac.jp

***Abstract*—When texts are mined for meaningful information, one important aspect is to construct a coding rule that categorizes key terms into several conceptual groups. Usually such a rule is human-made and tends to be subjective. The present study attempts to build coding rules automatically from the ISO 26000 document and compares the results with those obtained by creating the coding rules manually.**

I. Introduction

There are mainly two approaches to mining text data quantitatively. One is the data-driven approach that attempts to identify key features within the text data without using any previously created rules. The other is to use human-made rules to analyze or classify the text data.

The former approach allows us to examine the data objectively using some statistical methods including multivariate analysis. However, when there is a hypothesis or a particular perspective a researcher aims to use as basis for their investigation, this approach proves to be inflexible in that the statistically calculated features do not always correspond with what the researcher has in mind.

The latter approach uses criteria provided by a researcher for text analysis, which is generally known as a "dictionary-based" approach. In the dictionary, words are grouped into semantic categories that are created manually. For example, "angry," "happy," and "sad" could be categories of "emotional state," and this can be written as a coding rule "angry, happy, sad $\rightarrow$ emotional state."

Coding rules make it possible to analyze text data for specific purposes; if there were a need to examine the reputation of a product, for example, one could simply use coding rules such as "good, nice, great $\rightarrow$ positive" and "bad, poor, awful $\rightarrow$ negative". However, this would require much time and effort to establish valid rules especially when it is necessary to encode a wide range of concepts.

The present study overcomes these disadvantages by attempting to create coding rules automatically from a "base document," which provides conceptual descriptions of the target categories. This enables us to set up rules objectively on the basis of the data (base document) and at the same time to include specific perspectives from which the text data should be analyzed. The automation can also be expected to be helpful when creating a large number of rules, particularly if they are complex.

II. Related Work

Content analysis, which is a research method for analyzing texts, has been widely applied for the qualitative analysis of texts. The research presented in this paper can be regarded as one such approach in that it performs content analysis by constructing coding rules to be used for analyzing the text content.

In automated content analysis, a series of coding rules is created, and the texts are automatically categorized in accordance with the given coding rules. The conventional approach is to set up categories manually and then classify texts or words accordingly [1]; examples include survey studies for political texts [2] and a case study for classifying German online news [3]. Another possible approach is to classify target texts without predefined categories, including a method that identifies the categories by clustering texts [4] and one that uses categories extracted by employing the LDA (Latent Dirichlet Allocation) technique [5].

Other related studies include the construction of a dictionary by automatically extracting related words from each category in the coded texts [6]. Furthermore, another study carried out a thematic analysis of a dictionary [7].

It should be noted that the aim of the present study is not to suggest an automated method for conducting content analysis. Rather, it aims to demonstrate how to create coding rules automatically and examine the validity of this approach by comparing the results with those obtained by using manual coding rules.

III. Method for Extracting Coding Rules

Coding rules are created automatically by employing a keyword extraction method. The keyness of a word is quantified statistically by using SMART [8] based on TF and IDF.

978-1-4673-8985-3/16 $31.00 © 2016 IEEE
DOI 10.1109/IIAI-AAI.2016.210

If a document is organized according to some concepts, it is possible to extract keywords from each section (this type of document is referred to as a "base document" hereafter. ISO 26000 is one such example). This enables us to classify words into groups (e.g., keyword1, keyword2 → concept1), which can be used as coding rules to analyze other documents (e.g., CSR documents).

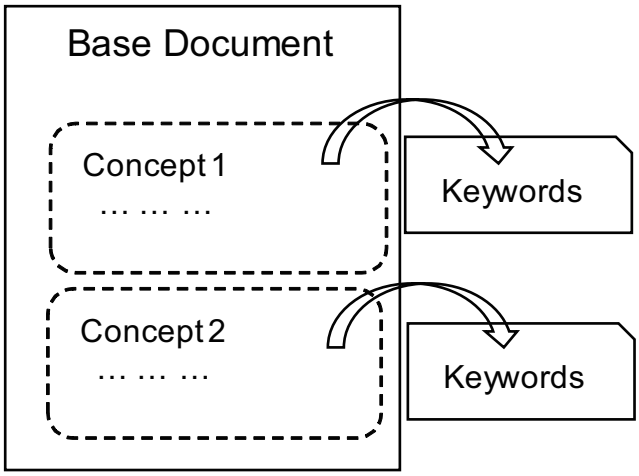

Fig. 1. Keyword Extraction from Each Section

IV. Experiment and Evaluation

A. Data

ISO 26000, the main target of this paper, is one of the most typical, popular, and accessible indexes of CSR (Corporate Social Responsibility). Although there are many definitions and conceptualizations of CSR, it basically signifies the active and voluntary mechanism that corporations use to engage stakeholders, including customers, employees, stockholders, consumers, and civil society as a whole in order to solve social and environmental problems. CSR has a number of indexes and guiding principles, such as human rights, labor practices, environmental protection, and anti-corruption policies. For instance, the Green Paper of the European Union briefly introduces the common features of CSR. Accordingly, it describes "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis" [9].

Although there are many CSR indexes, ISO 26000 plays a central role in corporate practices - especially in Japan. It has been argued that most Japanese corporations respect ISO 26000 and reiterate its principles in their own CSR reports. ISO 26000 is therefore a useful tool for analyzing corporate CSR documents in Japan. In terms of the conceptual documentation of ISO 26000, there is "*Guidance on social responsibility*" [1]. However, the Standards Documentation enumerates legal principles and, therefore, deviates slightly from the general expression of a company's CSR document. Another document, "*Comprehensive Social Responsibility: ISO 26000 and Cases of Small and Medium-Sized Business (Commentary)*" (hereafter referred to as Commentary), is published by the Japanese National Committee for the ISO Working Group on Social Responsibility (hereafter, National Committee). The Commentary is the document that explains the concept of the standard in plain terms; hence, we decided that it is suitable for the automated generation of coding rules. Therefore, this study uses this document as reference.

Commentary is a single, 18-page PDF file that can be downloaded from the National Committee's website[2]. ISO 26000 consists of seven guiding categories (organizational governance, human rights, labor practices, environment, fair operating practices, consumer issues, and community involvement and development).

B. Automated Generation of Coding Rules

This subsection explains how the coding rules were created from Commentary which was used as the base document. The third chapter of Commentary covers explanations of the categories listed in Table I (corresponding to sections 3.1-3.7). The automated coding process identified each paragraph in 3.1 to 3.7 as a different document.

A search engine was created using GETA [3] and only nouns (except for numerals) were extracted for the keyword analysis based on the results of the Japanese morphological analyzer ChaSen [4]. Then, these nouns were registered into WAM, which is the database of GETA. The keywords of each document were extracted on the basis of SMART, which is the default feature score of GETA, and the top 100 words were listed as keywords. Table I includes top 20 keywords for each category.

C. Analysis and Discussion

For the sake of comparison, one of our team members - a political science scholar and ISO 26000 expert - constructed coding rules manually for analyzing corporate CSR documents in accordance with Commentary. These manually coded rules appear in Table II, which lists the typical words in each category. These words are considered to be the correct set. We decomposed them into morphemes by using ChaSen to enable them to be compared with the automatically generated coding rules.

We obtained values for Precision and Recall by examining the top n keywords. Let A_{cn} be a set of keywords generated automatically for the category c, and let H_c be the set of correct words for the category c. Then, the values of Precision $\mathrm{P}(c,n)$ and Recall $\mathrm{R}(c,n)$ are calculated as follows:

$$\mathrm{P}(c,n) = \frac{|A_{cn} \cap H_c|}{|A_{cn}|},$$

$$\mathrm{R}(c,n) = \frac{|A_{cn} \cap H_c|}{|H_c|}.$$

[1] http://kikakurui.com/z26/Z26000-2012-01.html

[2] http://iso26000.jsa.or.jp/contents/,
Japanese National Committee for ISO Working Group on Social Responsibility (Japanese page only) (last accessed March 30, 2016).

[3] http://geta.ex.nii.ac.jp/e/,
Generic Engine for Transposable Association (GETA)

[4] https://en.osdn.jp/projects/chasen-legacy/ChaSen,
ChaSen: Morphological-analysis system

TABLE I. AUTOMATED CODING RULES FOR ISO 26000

Category	Illustration
Organizational governance	thing, stake, holder, other, medical, home, corporation, utilization, decision, intention, validity, governance, six, school, next, realization, state, specialty, many, command
Human rights	liberty, the weak, direct, human, avoidance, child, consideration, all, equality, body, violation, human right, right, confirmation, action, discrimination, situation, assistance, oversea, indirect
Labor practices	institution, office, condition, aged, international, minimum, goods, occupation, standard, both parties, Labor Standard Law, opportunity, member, employee, upbringing, negotiation, human resource, obligation, discipline, ability
Environment	use, place, thing, resource, diversity, certainty, disposal, substance, character, lower limit, -zation, effort, prevention, pollution, creature, climate, ecology, measure, discharge, change
Fair operating practices	promotion, chain, corporation, competition, injustice, subcontract, base, whole, fairness, project, right, profit, ethic, property, top, value, corruption, position, antitrust law, collusion
Consumer issues	product, system, management, individual, bad effect, use, service, strengthening, production, judge, method, information, consciousness, support, positive, introduction, supply, consumption, data, privacy,
Community involvement and development	health, object, technology, other, form, participation, cultivation, communication, culture, investment, community, local, involvement, skill, income, wealth, contribution, creation, development, tie

TABLE II. MANUAL CODING RULES FOR ISO 26000

Category	Illustration
Organizational governance	corporation, social responsibility, stakeholder, employee, audit, dialogue
Human rights	right, liberty, equality, human rights, discrimination, due diligence, complaint, overlook, the weak
Labor practices	labor, safety, health, ILO, employment, employee, social protection, social dialogue, human resource, work-life balance, irregular employment
Environment	environment, resource, pollution, climate change, ecosystem, development, prevention, energy saving, resource conservation, recycling, nature, sustainability, biodiversity, air, water, soil, purification, greenhouse gas
Fair operating practices	fair, operating practice, corruption, value chain, social responsibility, ethics, property right, consciousness, whistle-blowing, subcontractor, fair trade, customer, client
Consumer issues	safety, security, sanitation, defect, consumer, influence, quality, private information, data, marketing, contract, sustainability, complaint, dispute, privacy, consciousness, customer, eco
Community involvement and development	community, communication, health, social investment, job creation, shopping street, event, inhabitant, local economy, education, culture, technology, technique, volunteer, enlightenment, sports, homeless, involvement, development

The Precision-Recall curve is shown in Fig. 2.

The values of Precision and Recall have an opposite correlation. These values are required to be well balanced for obtaining information. The F_1score is useful for the general evaluation of Precision and Recall. The F_1score is the value calculated by the harmonic mean of Precision and Recall. The F_1score $\mathrm{F}_1(c,n)$ of the top n feature words is calculated by the following formula:

$$\mathrm{F}_1(c,n) = \frac{2 \cdot \mathrm{P}(c,n) \cdot \mathrm{R}(c,n)}{\mathrm{P}(c,n) + \mathrm{R}(c,n)}.$$

Fig. 3 shows the F_1score values of each category, and Table III shows the maximum of the F_1score values. The category with the highest value is "Community involvement and development," and its top 10 keywords are "**health**," "objective," "**technology**," "other," "form," "involvement," "development," "**communication**," "**culture**," and "**investment**." According to the manual coding rules (Table II), the correct coding should include the words shown in bold face.

On the other hand, "Organizational governance" has a low value. Among the top 34 keywords, only "stakeholder" is correctly coded. The reason for this is that the volume of the "Organizational governance" section in Commentary is small compared with other categories. For example, the sections of other categories in Commentary consist of five sub-sections, namely, the issues, remaining points for Japanese small and

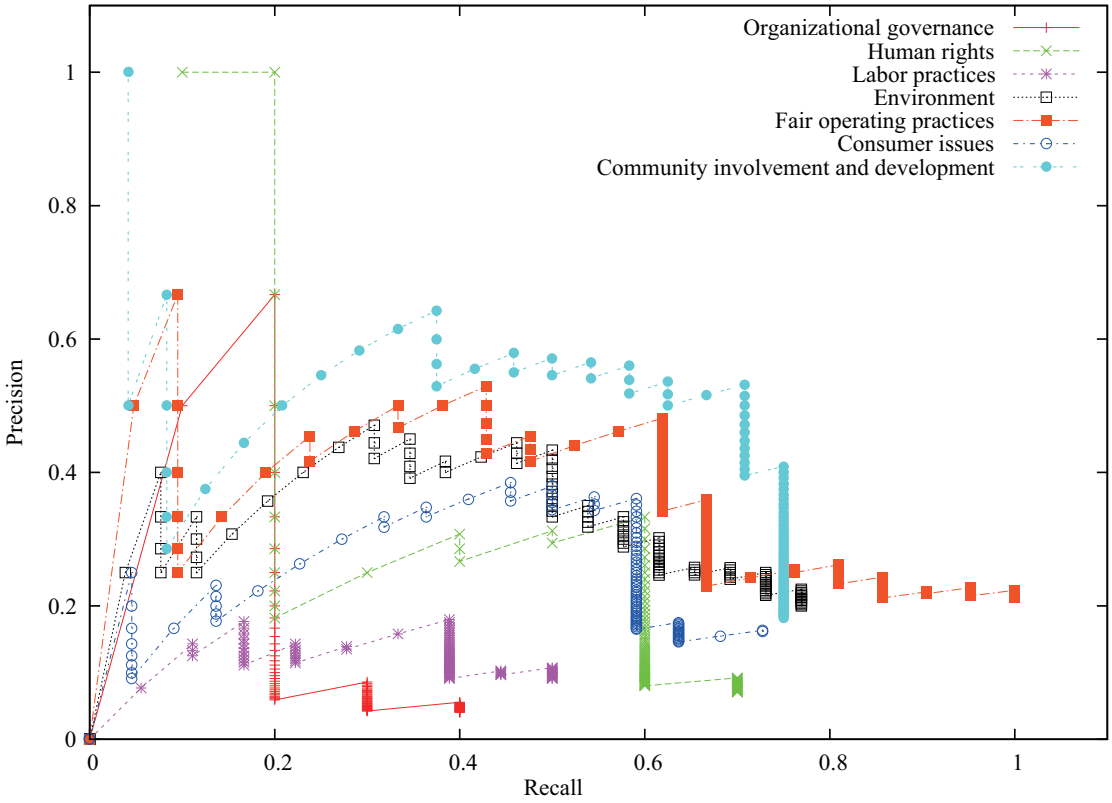

Fig. 2. Precision-Recall curve for the Automated Generation of Coding Rules

TABLE III. MAXIMUM OF F_1 SCORE OF EACH CATEGORY

Category	Maximum of F_1 score
Organizational governance	0.31
Human rights	0.43
Labor practices	0.25
Environment	0.46
Fair operating practices	0.54
Consumer issues	0.45
Community involvement and development	0.61

medium-sized business, concrete action plan, commentary to existing problems, and principal related laws, whereas the section of "Organizational governance" consists of only two sub-sections, namely, issues and the concrete action plan and principal related laws (only one page). Thus, the size of documents could affect the coding precision. Another point is that "labor practices" also shows a low value. The data for keywords indicate that no correct coding is found below the top 12 words. Further examination is necessary to analyze the mechanism and relationship between the size of documents and the coding precision.

V. CONCLUSION

The study presented in this paper shows the possibility of generating coding rules automatically using ISO 26000 as a base document. Key terms were extracted with respect to each category outlined in ISO 26000, which were turned into coding rules for each concept. Then, a comparison was made between the automated coding rules and manual coding rules, which were found to be in close agreement with each other. The present study remains at a preliminary stage in that some coding rules have low F_1score values, but these figures are expected to improve once the data size of the base document becomes larger. Specifically, the accuracy can be improved by using the ISO standard or other related documents. Another potential option would be to consider compounds and phrases as keywords. An additional future task would be to examine whether the automated coding rules are meaningful for mining useful information.

REFERENCES

[1] J. L. S. Yan, N. McCracken, and K. Crowston, "Semi-automatic content analysis of qualitative data," *iConference 2014 Proceedings*, 2014.

[2] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, 2013.

[3] M. Scharkow, "Thematic content analysis using supervised machine learning: An empirical evaluation using german online news," *Quality & Quantity*, vol. 47, no. 2, pp. 761–773, 2013.

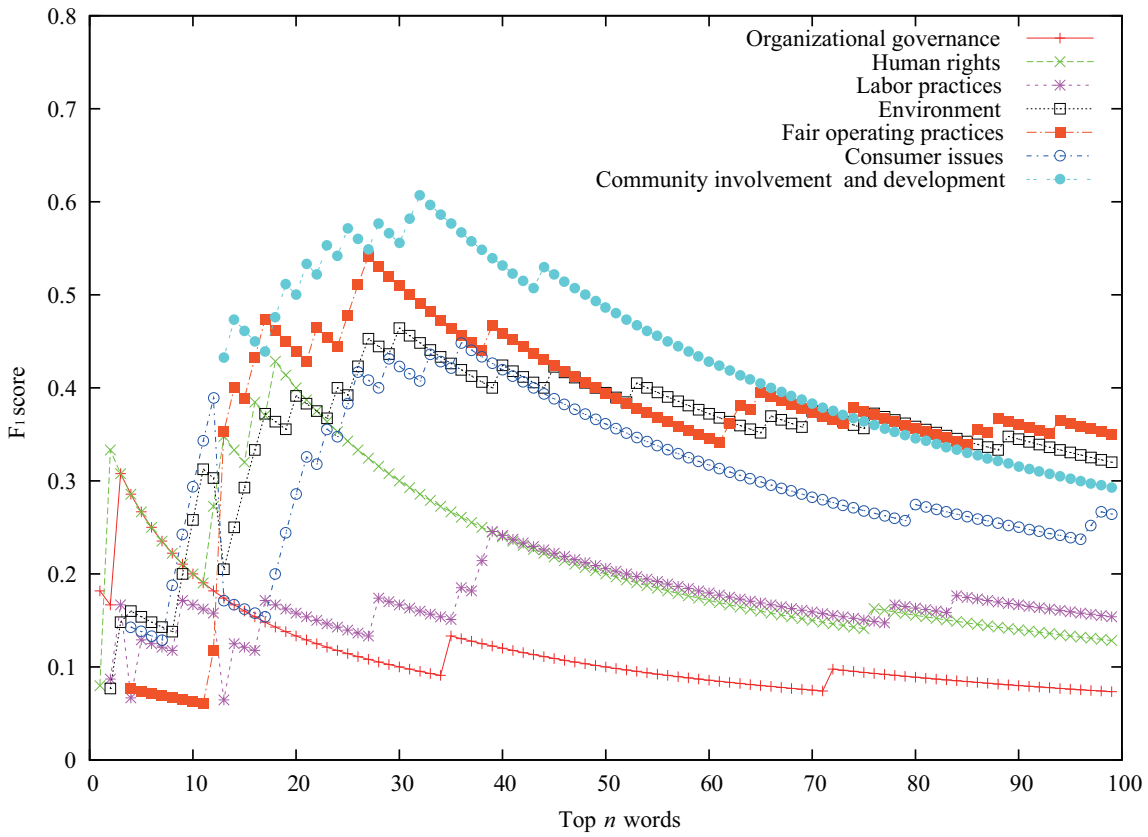

Fig. 3. F_1 scores for the Automated Generation of Coding Rules

[4] Y.-H. Chang, C.-Y. Chang, and Y.-H. Tseng, "Trends of science education research: An automatic content analysis," *Journal of Science Education and Technology*, vol. 19, no. 4, pp. 315–331, 2010.

[5] C. Zirn and H. Stuckenschmidt, "Multidimensional topic analysis in political texts," *Data & Knowledge Engineering*, vol. 90, pp. 38–53, 2014.

[6] Y. Takayama, Y. Tomiura, K. R. Fleischmann, A.-S. Cheng, D. W. Oard, and E. Ishita, "Automatic dictionary extraction and content analysis associated with human values," *Information Engineering Express*, vol. 1, no. 4, pp. 107–118, 2015.

[7] K. R. Fleischmann, Y. Takayama, A.-S. Cheng, Y. Tomiura, D. W. Oard, and E. Ishita, "Thematic analysis of words that invoke values in the net neutrality debate," *iConference 2015 Proceedings*, 2015.

[8] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 21–29.

[9] European Commission, *Green paper: promoting a European framework for corporate social responsibility*. Office for Official Publications of the European Communities, 2001.