International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2017, Vol.1, No.1, P.19 - 30

Performance Comparison on Automated Generation of Coding Rules: A Case Study on ISO 26000

Tetsuya Nakatoh[*] , Satoru Uchida[†] , Emi Ishita[‡] , Toru Oga[§]

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 by using two proposed methods. The results were compared with the manually created coding rules, and the SVM method was proven to be more effective.
Keywords: Text mining, Coding rules, Automated Generation, SVM, ISO 26000.

1 Introduction

Quantitative text data is mainly mined using two approaches: the data-driven approach that attempts to identify key features within the text data without using any previously created rules and the approach that uses human-made rules to analyze or classify the text data.

The data-driven approach allows us to examine the data objectively by using some statistical methods including multivariate analysis. However, when a hypothesis is formulated or researchers aim to adopt a particular perspective as a basis for their investigation, this approach proves to be inflexible because the statistically calculated features do not always correspond with what the researchers have in mind.

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

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

[*] Research Institute for Information Technology, Kyushu University, Fukuoka, Japan
[†] Faculty of Languages and Cultures, Kyushu University, Fukuoka, Japan
[‡] Research and Development Division, Kyushu University Library, Fukuoka, Japan
[§] Faculty of Law, Kyushu University, Fukuoka, Japan

The present study overcomes these disadvantages by attempting to automatically create coding rules 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 simultaneously include specific perspectives from which the text data could be analyzed. The automation can also be expected to be helpful when creating a large number of rules, particularly if they are complex.

We focused on the fact that the "base document" (see Section 4.1 for details), which provides a conceptual explanation of the target category, basically contains the information necessary for coding. We assumed that feature words extracted from documents can be used as coding rules. There are several methods of extracting feature words, for example TF-IDF and SMART used in our previous research [7]. In addition, methods described in [12, 8] are listed as feature selection methods for classification. In this paper, we used a simpler approach to obtain the importance of feature words directly from linear support vector machine (SVM) classifiers [1, 3]. We compared the performance of the proposed and SMART methods in the generation of coding rules.

2 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 the present paper can be regarded as one such approach as 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 according to the given coding rules. The conventional approach is to set up categories manually, and then classify texts or words accordingly [13]; examples include survey studies for political texts [6] and a case study for classifying German online news [9]. Another possible approach is to classify target texts without predefined categories; this includes a method that identifies the categories by clustering texts [2] and a method that uses categories extracted by employing the latent Dirichlet allocation technique [14].

Other related studies include the construction of a dictionary by automatically extracting related words from each category in the coded texts [11]. Another study performed a thematic analysis of a dictionary [5].

Note that the aim of the present study is not to suggest an automated method for conducting content analysis but to demonstrate the creation of coding rules automatically and examine the validity of this approach by comparing the results with those obtained using manual coding rules.

3 Method for Extracting Coding Rules

3.1 Analysis Target

In this paper, we analyzed not only nouns but also compound nouns because a coding rule generally contains compound nouns. The base document is divided into words with parts of speech by using ChaSen [1]. A continuous noun is connected as a compound noun. However, a part of compound noun may be important as a feature word. Therefore, we also retained nouns and composite nouns, which are part of compound nouns. Next, the numbers of

[1] Japanese Morphological-analysis system: `https://en.osdn.jp/projects/chasen-legacy/ChaSen`

occurrences of nouns and compound nouns were counted for each document, and are the target data in this analysis.

3.2 Feature word extraction using SMART

In our previous study [7], we used the method of normalizing word frequencies based on the document length. This method was proposed by Singhal et al. [10] and is used as a standard measure of the Generic Engine for Transposable Association (GETA) [2]. Hereafter, we refer to this measure as "SMART." We constructed a search engine by registering nouns and compound nouns contained in the documents of a word article matrix: a database system of the GETA. Then, we retrieved the feature words of the document by searching the document of the corresponding category as a key. In this approach, a number of compound words were listed as keywords for coding rules. Therefore, in our previous research, compound words were broken down into words to allow a precise comparison between another set of feature words. However, to achieve a more rigorous evaluation, it is necessary to evaluate compound words directly. Therefore, in the current study, we modified the algorithm so that it can manage compound words.

3.3 Feature word extraction using SVM

SVM is one of the pattern recognition models using supervised learning. Given a set of training examples, an SVM training algorithm builds a binary linear classifier that assigns new examples to one category. In linear classification, it is possible to obtain weights for each attribute of feature vectors used for training from the classifier constructed by the SVM. In this paper, we used the attribute weight as the feature value of a word to select the candidates for coding rules. Specifically, we constructed a classifier by using an SVM that distinguishes documents of a specified category from other categories. We adopted nouns and compound nouns as attributes for the classification. Each category comprises a feature vector based on the number of occurrences of the attributes. SVM generates a classifier that distinguishes a category from the other by using their feature vectors as input. After that, the feature values of each feature word can be calculated from the classifier obtained through SVM learning. The feature values are the sum of the product of the coefficients of each support vector and the weight of each attribute in the support vector included in the classifier.

4 Experiment and Evaluation

4.1 Data

ISO 26000, the main target of this paper, is one of the most typical, popular, and accessible indexes of corporate social responsibility (CSR). Although there are many definitions and conceptualizations of CSR, it basically signifies the active and voluntary mechanism used by corporations to engage stakeholders, including customers, employees, stockholders, consumers, and the civil society as a whole, to solve social and environmental problems. CSR has a number of indexes and guiding principles, such as human rights, labor practices, environmental protection, and anticorruption policies. For instance, the Green Paper of

[2] Generic engine for transposable association: `http://geta.ex.nii.ac.jp/e/`

the European Union briefly introduces the common features of CSR. Accordingly, it describes it as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis" [4].

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 follow ISO 26000 and reiterate its principles in their own CSR reports. Therefore, in Japan, ISO 26000 is a useful tool for analyzing corporate CSR documents. The conceptual documentation of ISO 26000 is described in the "*Guidance on social responsibility*" [3]. However, the standards documentation enumerates legal principles, and therefore deviates slightly from the general expression of a company's CSR document. Another document, called the "*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 referred to as National Committee). The Commentary is a document that explains the concept of the standards in plain terms; hence, we decided that it is suitable for the automated generation of coding rules, and used it as a reference in the current study.

The Commentary is a single, 18-page PDF file that can be downloaded from the National Committee's website[4]. ISO 26000 consists of seven guiding categories: organizational governance, human rights, labor practices, environment, fair operating practices, consumer issues, and community involvement and development.

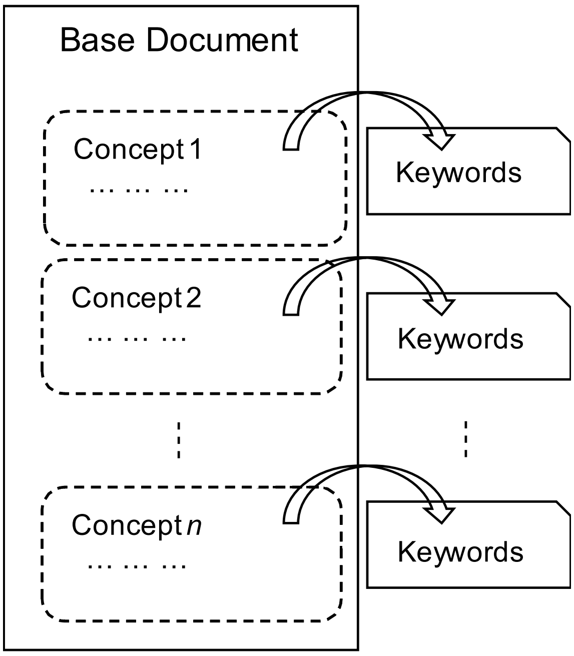

Figure 1: Keyword Extraction from Each Section

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

[4] Japanese National Committee for ISO Working Group on social responsibility (Japanese page only; last accessed March 30, 2016): `http://iso26000.jsa.or.jp/contents/`

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). Figure 1 illustrates the concept.

4.2 Correct Set of Coding Rules

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 the Commentary. These manually coded rules are listed in Table 1, which lists the typical words in each category. These words are considered to be the correct set. To construct coding rules that represent each category precisely, there is no limit for the number of typical words. In addition, typical words include phrases.

Table 1: 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

4.3 Automated Generation of Coding Rules

By using SMART and SVM, the coding rules were created from the Commentary, which was used as the base document. The third chapter of the Commentary covers explanations of the categories listed in Tables 1, 2, and 3 (corresponding to sections 3.1–3.7). The automated coding process identified each paragraph in 3.1 to 3.7 as a different document.

Table 2: Automated Coding Rules based on SMART

Category	Illustration
Organizational governance	object, stake, **stakeholder**, holder, company, medical, house, utilization, other, decision, intention, decision making, many, school, realization, specialty, next, state, organizational governance, leadership
Human rights	direct, child, avoidance, cultural, economic, **freedom**, the weak, **the weak**, man, equality, basic principle, body, physical, infringement, consideration, all, **human rights**, **discrimination**, human rights violations, due diligence
Labor practices	**ILO**, workplace, conditions, system, labor law, both, product, international, minimum, level, profession, old age, senior citizens, **employee**, opportunity, member, employment, worker, **human resources**, human resource development
Environment	place, object, variety, use, **resource**, processing, material, certainty, sex, minimum, **prevention**, variety of organisms, **biodiversity**, prophylactic, fluctuation, countermeasure, discharge, climate, **climate change**, environment issues
Fair operating practices	chain, ahead, injustice, competition, promotion, subcontract, subcontract act, overall, foundation, **fair**, **social responsibility**, business, right, **ethics**, ethical, **operating practice**, property, **property right**, top, value
Consumer issues	system, management, an individual, use, adverse effect, awareness raising, product, service, a judgment, reinforcement, method, positive, manufacture, information, **consciousness**, introduction, support, positive, consumption, consumer problem
Community involvement and development	**health**, object, **technology**, other, join, form, development of, **communication**, **community**, region, contribution, creation, **development**, **involvement**, **inhabitant**, local resident, wealth, income, **technique**, skill development

The top 100 feature words for each category were obtained based on SMART. Table 2 contains the top 20 feature words of each category. Similarly, top 100 feature words for each category were obtained using SVM. Table 3 contains the top 20 feature words of each category obtained using SVM.

Table 3: Automated Coding Rules obtained through SVM

Category	Illustration
Organizational governance	governance, **stakeholder**, holder, stake, **corporation**, decision making, intention, decision, behavior, object, company, **dialogue**, structure, other, offer, concrete, concrete, concrete behavior, **dialogue**, core
Human rights	**human rights**, infringement, right, **discrimination**, human, target, human rights violations, **due diligence**, situation, basic, **overlook**, socially vulnerable, **the weak**, respect, action, direct, indirect, basic right, child labor, citizen
Labor practices	**labor**, worker, human, **employment**, conditions, **employee**, **safety**, working conditions, discussions, human resource development, **human resources**, training, labor practice, a member, employment, workplace, **ILO**, relationship, opportunity, sanitation
Environment	**environment**, **resource**, influence, variety, use, **prevention**, **pollution**, creatures, sex, efforts, **biodiversity**, various organisms, **climate change**, climate, discharge, ecology, fluctuation, prophylactic, countermeasure, environment issues
Fair operating practices	**fair**, **ethics**, ethical, business, property, **operating practice**, responsibility, competition, injustice, practice, **social responsibility**, society, prevention, promotion, right, involvement, transaction, **property right**, politics, profit
Consumer issues	consumption, **consumer**, a person, service, information, product, consumer problem, **safety**, **consciousness**, task, an individual, **data**, protection, offer, system, management, **contact**, proper, **private information**, necessary
Community involvement and development	**community**, region, contribution, **technology**, creation, **education**, development of, development, employment, **involvement**, residents, **health**, **culture**, improvement, local resident, **technique**, wealth, income, investment, object

5 Analysis and Discussion

Feature words obtained using the two methods were used as candidates for the elements of the coding rule. These lists were evaluated based on Table 1 as the correct answers.

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

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

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

The Precision–Recall curves obtained using SMART and SVM are shown in Figures 2 and 3, respectively. Although differences occur depending on categories, it is possible to observe the clear difference in precision from these two graphs. In [7], feature word extraction based on SMART was effective to some extent; however, as pointed earlier, compound nouns were not handled well, and this is a serious drawback of the approach. In contrast, SVM is able to extract feature words including compound nouns relatively well.

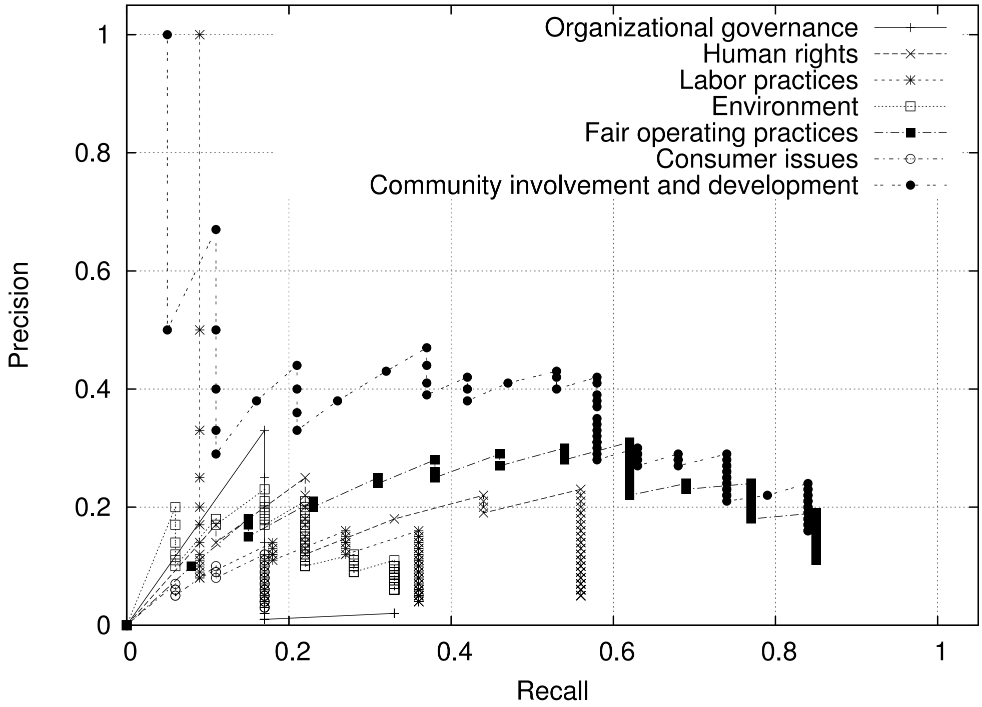

Figure 2: Precision–Recall curve for Automated Coding Rules by using SMART

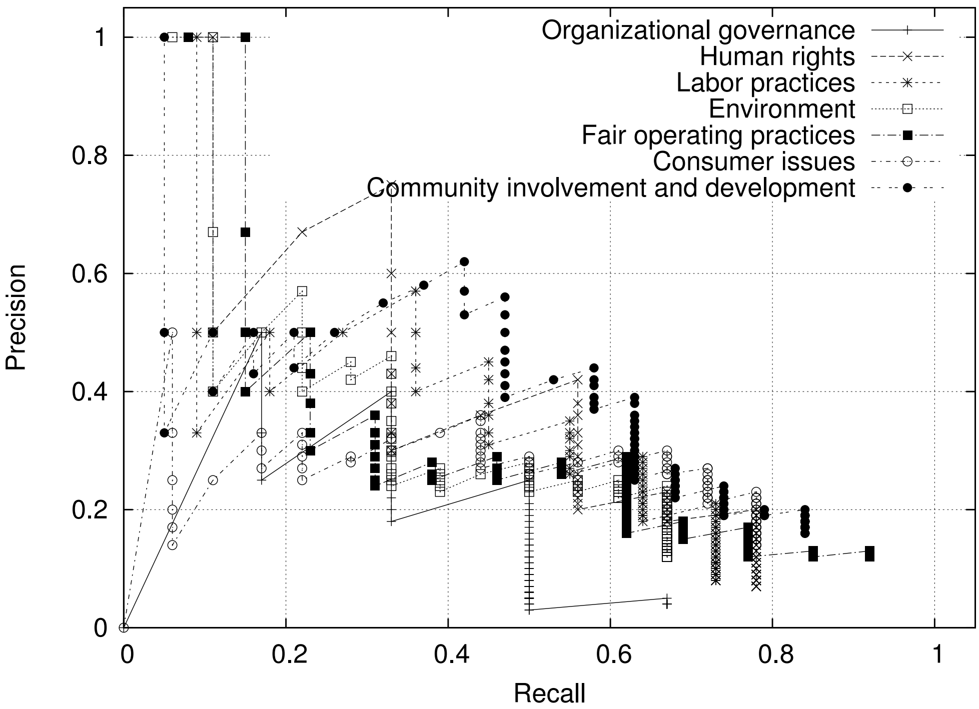

Figure 3: Precision–Recall curve for the Automated Coding Rules by using SVM

The values of Precision and Recall have an opposing correlation, and are required to be well balanced for obtaining information. The F_1score is useful for the general evaluation of Precision and Recall, and is the value calculated by the harmonic mean of Precision and Recall. The F_1score $F_1(c,n)$ of top n feature words is calculated as follows:

$$F_1(c,n) = \frac{2 \cdot P(c,n) \cdot R(c,n)}{P(c,n) + R(c,n)}.$$

Figures 4 and 5 show the Precision, Recall, and F_1score of the extraction results that summarize all the categories for the two methods. The value of Recall on SVM increases with the value of n, and the value of Precision on SVM is relatively large when the value of n is small. Therefore, it can be seen that the method using SVM is more suitable for the automated generation of coding rules than the method using SMART.

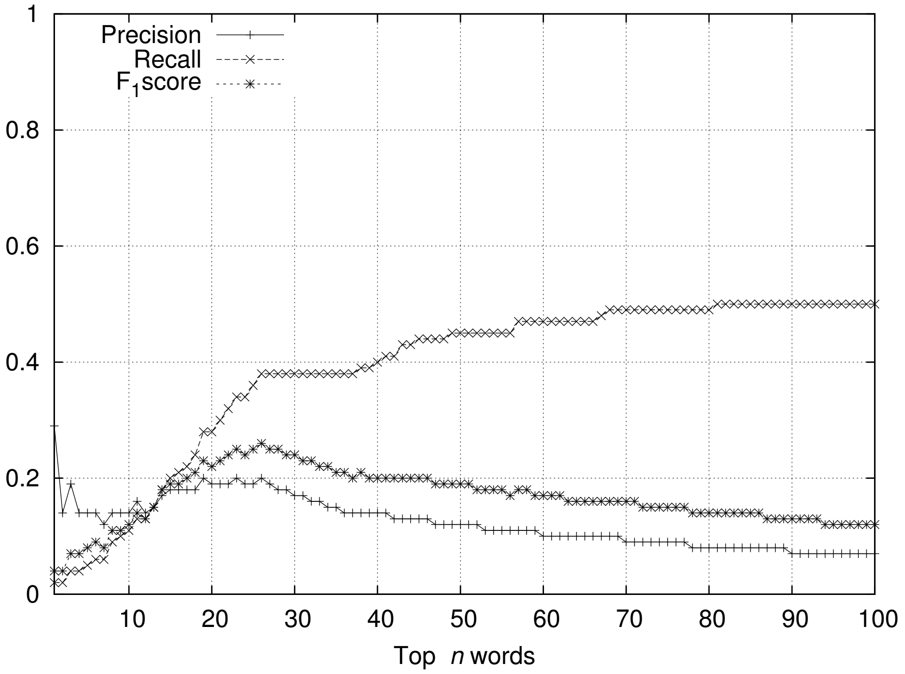

Figure 4: Precision, Recall, and F_1scores for the automated coding rules obtained through SMART

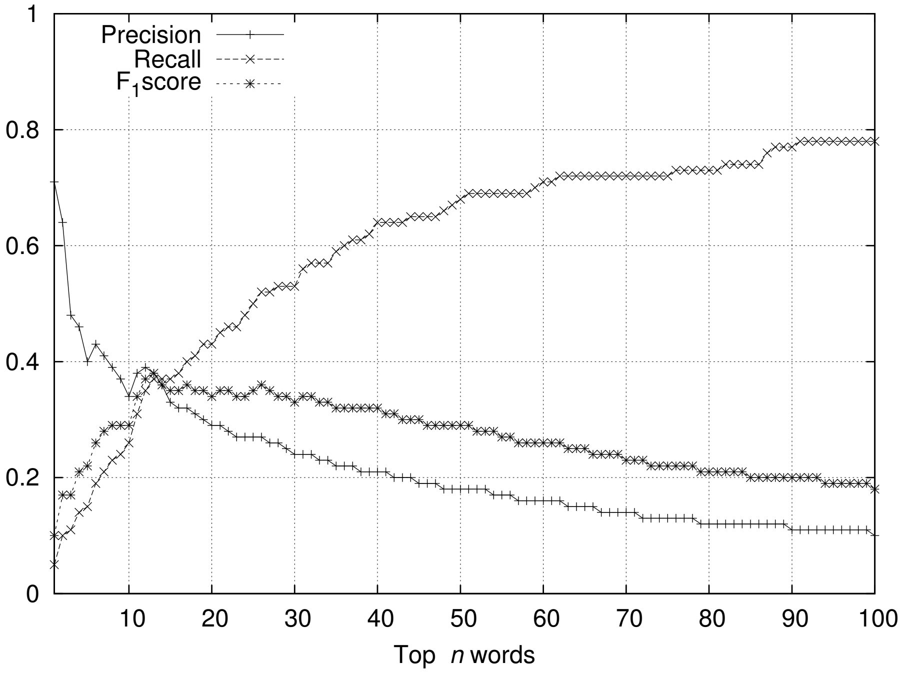

Figure 5: Precision, Recall, and F_1scores for the automated coding rules obtained through SVM

Table 4 indicates the maximum values of F_1score in each method. It is observed that the scores are generally better for SVM. The only exception is the category "Fair operating practices," in which SMART yields a better score but with a minor difference. In addition, it can be pointed out that while the scores of SVM are stable in any category, the scores of the SMART method are unstable and low overall.

Table 4: Maximum of F_1score of each category

Category	Maximum of F_1score	
	SMART	SVM
Organizational governance	0.22	0.36
Human rights	0.32	0.48
Labor practices	0.22	0.45
Environment	0.22	0.39
Fair operating practices	0.41	0.39
Consumer issues	0.14	0.41
Community involvement and development	0.49	0.51

Feature words in Tables 2 and 3 are shown in bold when they match the coding rule in Table 1. It should be noted that compound words may require special attention. For example, in Table 3, "climate change" is one of the correct answers in the environment category. Tables 1 and 2 appropriately list this compound noun as a coding rule but both tables include "climate," which is actually a part of the compound noun. To avoid subjective judgments, the present study did not make any special adjustments with respect to compound nouns. However, in our future research, we plan to reduce the effect of compounds by subtracting the number of compound nouns from the noun counts, or by normalizing the word frequencies based on word length.

6 Conclusion

This paper presented the possibility of generating coding rules automatically with the use of ISO 26000 as the base document. In this study, key terms were extracted with respect to each category outlined in ISO 26000, and turned into coding rules for each concept. Then, the manual coding rules and were compared with the two types of automated coding rules using SMART and SVM. The results clearly indicated that SVM outperforms SMART in the generation of coding rules and is good at treating compound nouns.

In this analysis, we simply evaluated the correspondences between the expert and machine coding rules. However, some feature words were not included in the expert list but still seem to represent the category, for example "decision making" and "sanitation" in Organizational governance and Labor practices respectively. Although a further examination is necessary in this respect, it seems that the automated generation of coding rules may support experts in creating their own coding rules. An additional future task would be to examine whether the automated coding rules are meaningful for mining useful information.

References

[1] J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic. Feature selection using support vector machines. *WIT Transactions on Information and Communication Technologies*, 28, 2002.

[2] 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*, 19(4):315–331, 2010.

[3] Y.-W. Chang and C.-J. Lin. Feature ranking using linear svm. In *WCCI causation and prediction challenge*, pages 53–64, 2008.

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

[5] 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.

[6] J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

[7] T. Nakatoh, S. Uchida, E. Ishita, and T. Oga. Automated generation of coding rules: Text-mining approach to ISO 26000. *5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI2016)*, pages 154–158, 2016.

[8] M. H. Nguyen and F. De la Torre. Optimal feature selection for support vector machines. *Pattern recognition*, 43(3):584–591, 2010.

[9] M. Scharkow. Thematic content analysis using supervised machine learning: An empirical evaluation using german online news. *Quality & Quantity*, 47(2):761–773, 2013.

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

[11] 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*, 1(4):107–118, 2015.

[12] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 647–653. MIT Press, 2000.

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

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