JAPAN EXCHANGE GROUP

JPX
WORKING
PAPER

Examination of Machine Translation in Corporate Governance Reports

Nobushige DOI, Masafumi KONDO, Atsushi SANTO

May 8, 2019

Vol. 31

Remarks

The JPX Working Paper is a compilation of research and research results by executives and employees of Japan Exchange Group, Inc. and its subsidiaries and affiliates (hereinafter referred to as "Japan Exchange Group, etc."), and external researchers. It is intended to receive a wide range of comments from academic societies, research institutes, market participants and others. The contents and opinions of the published papers belong to the individual author and do not represent the official views of Japan Exchange Group, etc.

Examination of Machine Translation in Corporate Governance Reports[*][†]

Nobushige DOI[‡] Masafumi KONDO[§] Atsushi SANTO[¶]

2019/05/08

Abstract

The presence of overseas investors is increasing in the Japanese securities market. However, only about 4.4% of listed companies disclose their CG Reports in English, despite them being one of the important information sources in investment decisions.

In this study, we sorted the problems in machine translation of CG Reports and built a parallel corpus using existing CG Reports. Then, we created a custom translation model that applied domain adaptation to a general NMT model and evaluated the translation quality to verify the possibility of adopting machine translation for CG Reports.

Our experiments confirmed that BLEU scores increased by about 6 points when English translations of technical terms, proper nouns, and other expressions were modified.

---

[*] This paper is a modification and revision of "コーポレート・ガバナンス報告書における機械翻訳の検討 (Consideration of machine translation in corporate governance reports)"[1] published in Proceedings of the 25th Annual Conference of the Association for Speech Processing (NLP2019) (March 2019).

[†] In preparing this paper, we received helpful comments from staff members of Japan Exchange Group, etc. We are deeply grateful to them.

[‡] Fintech Laboratory, Strategy Planning Department, Japan Exchange Group, Inc. (n-doi [at] jpx.co.jp)

[§] Senior Business Analyst, Fintech Laboratory, Strategy Planning Department, Japan Exchange Group, Inc. (m-kondo [at] jpx.co.jp), Certificate-holder Member of the Securities Analysts Association of Japan

[¶] Director, Fintech Laboratory, Strategy Planning Department, Japan Exchange Group, Inc. (a-santo [at] jpx.co.jp), Certificate-holder Member of the Securities Analysts Association of Japan

1 Introduction

The presence of overseas investors is increasing in the Japanese securities market. According to the Shareownership[2] and Trading by Type of Investors[3] surveys published by Tokyo Stock Exchange, Inc. ("TSE"), the ratio of share ownership of listed companies in Japan by foreign corporations (based on market capitalization) has risen almost consistently since around 1990, and is about 30% according to the former survey as of the end of March 2018. Also, the ratio of overseas investors' transactions to share trading value is similar at approximately 58% in 2017.

One of the important sources of information on investors' investment decisions is the "Corporate Governance[*1] Report" ("CG Report(s)") disclosed by each listed company[4]. Under the Securities Listing Regulations, TSE requires listed companies to submit to TSE the status of compliance with each principle in Japan's Corporate Governance Code[*2] ("CG Code") in the CG Report[*3].

Listed companies describe their status of corporate governance in CG Reports in a format that can be analyzed by investors. In addition to Japanese CG Reports, English CG reports can also be disclosed. According to TSE, "disclosure of the CG report in English is desirable, particularly for companies with a high percentage of foreign shareholders."[5]. However, the proportion of TSE-listed companies that disclose English CG Reports is still approximately 4.4% (as of the end of March 2018).

Based on these circumstances, we propose the use of machine translation as a way to make it easier to disclose English CG Reports to overseas investors. Machine translation that employs a neural machine translation model[6] ("NMT model") is known to produce higher quality translations than those that employ conventional machine translation models. However, having source text that contains many domain-specific terminologies results in lower quality translations for an NMT model where training data sets are based on general-purpose algorithms[7]. Furthermore, there are limited language resources that can be used in training data sets for NMT models tailored to CG Reports. One solution to these problems is a method called Domain Adaptation, which improves the quality of machine translation for a specific field by combining general-purpose algorithms and training data of a specific domain. The domain-adapted machine translation model is called a "Custom Translation Model" or "Custom Model"[8, 9].

In this study, after sorting problems in machine translation of CG Reports, we built a parallel corpus from text in CG Reports already disclosed in both Japanese and English to create a domain-adapted NMT model. Then, we evaluated the quality of translations and examined the possibility of using machine translation for CG Reports. The results of our experiments confirmed that BLEU scores[10] were higher when creating a custom translation model with the constructed parallel corpus compared to a general-purpose NMT model.

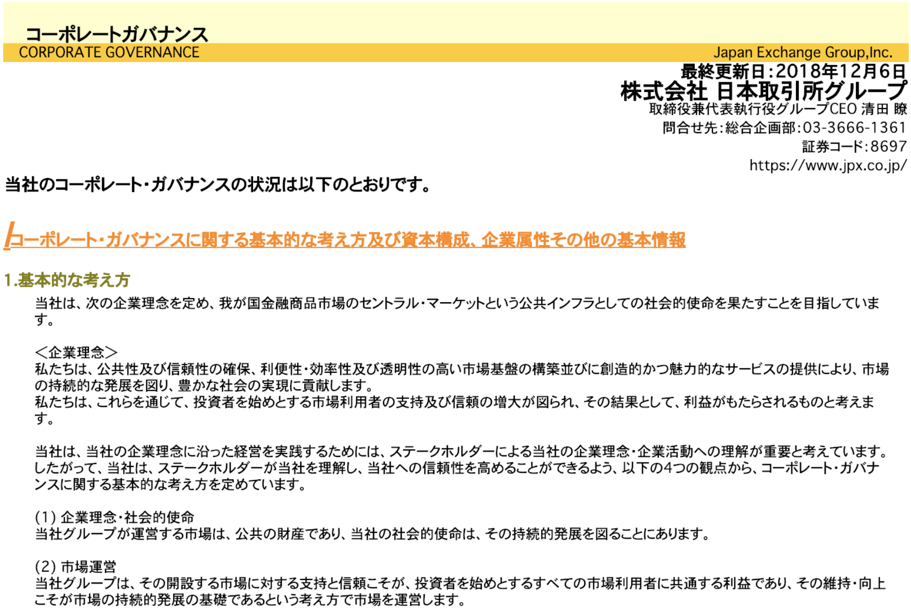
[*1] The CG Code defines "corporate governance" (CG) as "[a] structure for transparent, fair, timely, and decisive decision-making by companies, with due attention to the needs and perspectives of shareholders and also customers, employees and local communities."

[*2] The Corporate Governance Code is a code that summarizes major principles contributing to the realization of effective corporate governance and is positioned as an attachment to the Securities Listing Regulations.

[*3] See Articles 419 and 436-3 of the Securities Listing Regulations.

Table. 1　Main Sections in CG Reports

I	Basic Views on Corporate Governance, Capital Structure, Corporate Attributes and Other Basic Information
II	Business Management Organization and Other Corporate Governance Systems regarding Decision Making, Execution of Business, and Oversight
III	Implementation of Measures for Shareholders and Other stakeholders
IV	Matters Related to the Internal Control System
V	Other



コーポレートガバナンス
CORPORATE GOVERNANCE　Japan Exchange Group,Inc.

最終更新日:2018年12月6日
株式会社 日本取引所グループ
取締役兼代表執行役グループCEO 清田 瞭
問合せ先:総合企画部:03-3666-1361
証券コード:8697
https://www.jpx.co.jp/

当社のコーポレート・ガバナンスの状況は以下のとおりです。

Iコーポレート・ガバナンスに関する基本的な考え方及び資本構成、企業属性その他の基本情報

1.基本的な考え方

当社は、次の企業理念を定め、我が国金融商品市場のセントラル・マーケットという公共インフラとしての社会的使命を果たすことを目指しています。

＜企業理念＞
私たちは、公共性及び信頼性の確保、利便性・効率性及び透明性の高い市場基盤の構築並びに創造的かつ魅力的なサービスの提供により、市場の持続的な発展を図り、豊かな社会の実現に貢献します。
私たちは、これらを通じて、投資者を始めとする市場利用者の支持及び信頼の増大が図られ、その結果として、利益がもたらされるものと考えます。

当社は、当社の企業理念に沿った経営を実践するためには、ステークホルダーによる当社の企業理念・企業活動への理解が重要と考えています。したがって、当社は、ステークホルダーが当社を理解し、当社への信頼性を高めることができるよう、以下の4つの観点から、コーポレート・ガバナンスに関する基本的な考え方を定めています。

(1) 企業理念・社会的使命
当社グループが運営する市場は、公共の財産であり、当社の社会的使命は、その持続的発展を図ることにあります。

(2) 市場運営
当社グループは、その開設する市場に対する支持と信頼こそが、投資者を始めとするすべての市場利用者に共通する利益であり、その維持・向上こそが市場の持続的発展の基礎であるという考え方で市場を運営します。

Fig. 1　Sample CG Report in Japanese

2　CG Reports

TSE prescribes a specific format and instructions on how to prepare the CG Reports disclosed by listed companies. The five main sections of the CG Report are listed in Table 1. An example of a Japanese CG Report is shown in Fig. 1. Materials disclosed by listed companies via TSE are usually in Portable Document Format (PDF). However, for Japanese CG Reports, listed companies are required to submit XBRL[*4] documents along with the PDF files. CG Reports in XBRL format contain tags for each section of the document, facilitating mechanical processing as compared to PDF files.

3　Characteristics of Japanese in CG Reports

As a preliminary experiment of this study, Japanese sentences in CG Reports were input into multiple machine translation services, and trends of mistranslations and features of Japanese sentences that are prone to mistranslations were considered. The tendencies considered to be problems in machine translation of CG Reports, which were noted in this preliminary experiment, are presented below.

[*4] An XML-based markup language used in accounting documents, including financial statements.

3.1 Technical terms and proper nouns

CG Reports contain many proper nouns, including company names and technical terms in the fields of Investor Relations (“IR”) and securities. It is often difficult to translate these correctly with general-purpose machine translation services. In addition, some technical terms may not have unique English translations. Examples are shown in Table 2.

Table. 2 Examples of Technical Terms and Proper Nouns

Source Text (Japanese)	Reference Translations[*5]	Machine Translation Service Output
監査役	Audit & Supervisory Board Member	Auditor
	Audit and Supervisory Board Members	
	Statutory auditor	
	Kansayaku	
議決権電子行使プラットフォーム	Electronic Voting Platform	Voting right electronic exercise platform
	Platform for the electronic exercise of voting rights	
EY 新日本有限責任監査法人	Ernst & Young ShinNihon LLC	EY New Japan limited liability audit corporation
PwC あらた有限責任監査法人	PricewaterhouseCoopers Aarata LLC	PwC Yuru limited liability audit corporation

3.2 Pronouns with Multiple Translation Variants

In many CG Reports, words such as “当社 (our company)” are used as pronouns referring to the company writing the document and “同氏 (said person)“ as a pronoun pointing to a previously mentioned person. There are multiple candidate English translations for such pronouns, and it is necessary to understand the context and standardize the writing style[*6]. Examples are shown in Table 3.

3.3 Coordinating conjunctive phrases in a single sentence

Listed companies often section off and use lists when describing their corporate governance initiatives. This results in CG Reports with bulleted or numbered lists and sentences that are broken down into coordinating conjunctive phrases. In particular, if a sentence containing bullets with line breaks is input into a

---

[*5] Reference translations in this paper are examples of actual phrasing found in English CG Reports.

[*6] Refers to usage of okurigana, parentheses, symbols, and other elements in Japanese, as well as structure or word choice[11]. Generally, documents that establish style requirements and standards are called “style guides”.

Table. 3 Examples of Pronouns with Multiple Translation Variants

Pronoun (Japanese)	Model Translation (English)	Reference Translations in Context from CG Reports
当社	We	We fully comply all aspects of the Corporate Governance Code.
	The Company	The Company implements each principle of the Corporate Governance Code.
	(Company name)	(Company name) implements each of the basic principles of the Corporate Governance Code.
同氏	He / She	He has never been the Company's legal advisor.
	(personal name)	(personal name) has extensive knowledge and experience as a lawyer.

machine translation service that translates line by line, mistranslations tend to occur. Examples of this are shown in Table 4. In addition, such sentence structure also causes problems when matching components during construction of parallel corpora.

Table. 4 Examples of coordinating conjunctive phrases in a single sentence

Source Text (Japanese)	Reference Translations	Machine Translation Service Output
強化すべき点としましては、当社事業がグローバルに拡大する中において、ガバナンス機能の更なる充実に向けた取組みが重要との認識に立ち、	Recognizing that efforts for further enhancement of the governance function is vital to Santen in the midst of the global expansion of its business, Santen has decided to exert further efforts with respect to items that need to be strengthened including those listed below:	As a point of strengthening, we recognize that efforts aimed at further enhancement of the security function are important as our business expands to global.
・監督機能をより強化するため、取締役会で意思決定した重要な事項に対する継続的なモニタリングをより充実させること	With the aim of strengthening the monitoring function, further enhancing the continuous monitoring of material matters that are decided at meetings of the Board of Directors; and	・ Enhance continuous monitoring of important matters made by the Board of Directors in order to enhance the supervision function.
・リスクマネジメント視点での議論を更に強化すること	Further strengthening discussions from the viewpoint of risk management.	・ To further strengthen the discussion from the perspective of risk management,
などについて、取り組むこととしました。		I decided to work on what.

3.4 Inexhaustive noun series

Context and previous knowledge of the term(s) used may be necessary to correctly translate a sentence containing words or phrases that indicate that the noun(s) is just one example or instance in an Inexhaustive series, such as the use of "等 (etc)." in "甲, 乙, 丙等 (A, B, C, etc.)." An example of this is shown in Table 5.

Table. 5 Example of an inexhaustive noun series

Source Text (Japanese)	Reference Translations	Machine Translation Service Output
最高経営責任者等の後継者の計画	Succession planning for positions such as president	Planning for successors of chief executive officers

3.5 Symbols for ordered lists

Various symbols are used for ordered lists in CG Reports. These are classified into two types: half-width alphanumeric characters “(1) (2) (3)” or “a) b) c)” and characters that require conversion to English, such as circled numbers, full-width Chinese-style numbers, or katakana letters. Particularly, in the case of the katakana letter “(イ)”, it can be used to point to the second “(ア)(イ) (ウ)” or first “(イ) (ロ) (ハ)” item in a list depending on the order in Japanese, which should be taken into consideration when translating it to English.

3.6 Coined words and poetic expressions

The CG Report has a section where companies describe such matters as their mission statement. There is a tendency for this section of CG Reports to contain coined words and poetic expressions that are derived from nuances unique to the Japanese language. When these sentences are input into a machine translation service, those nuances may be lost and the meaning may not be transmitted correctly. Examples of this are shown in Table 6.

Table. 6 Examples of coined words and poetic expressions

Source Text (Japanese)	Reference Translations	Machine Translation Service Output
当社は、経営理念 (Mission) を「日本を世界一豊かに。その未来へ心を尽くす一期一会の『いちご』」とし、定款に定めております。	Ichigo’s mission is to preserve and improve real estate and contribute to a sustainable society.	We have defined our management philosophy (Mission) as “the world’s richest in the world,” “Ichigo in a once-in-a-lifetime meeting to the future”, stipulated in the Articles of Incorporation.
挑戦することを真剣に楽しみ、独創的な“道 (どう)”を極め続けます。	Embrace challenges and seek to master the Doh (“Way” or “Path”) of creativity.	I am seriously enjoying challenging and continuing to keep my ingenious “way”.

4 Parallel Corpus of CG Reports

We built a parallel corpus by manually aligning the CG Code and CG Reports disclosed in both Japanese and English, and then pre-processing them. The process flow is shown in Fig. 2.

The following documents were used in the aforementioned process: first editions of the CG Code in

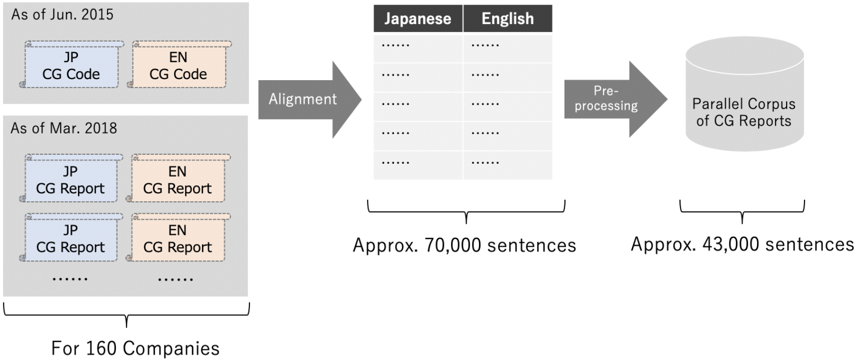


Fig. 2 Process for building parallel corpus

Japanese and English published in June 2015 (approximately 40 A4-size pages in total, each 210 × 297 cm) and the latest CG Reports of 160 companies disclosed both in Japanese and English as of March 2018.

Alignment in this paper refers to the task of creating parallel text by assigning the corresponding English sentence to a certain Japanese sentence. In this study, line breaks and punctuation marks at the ends of clauses are regarded as sentence breaks. Given that not all Japanese and English sentences in the documents necessarily correspond directly to each other on a per sentence basis, there were also cases where bilingual sentence sets had more than one sentence for either the Japanese or English text. After alignment, approximately 70,000 sentences were eligible for comparison.

Tasks including those listed below were carried out in pre-processing to ensure consistency of the meaning of parallel texts[*7]. In this process, we omitted parallel text that contained more than one sentence in Japanese and parallel text whose meaning is difficult to match in Japanese and English, such as in the example given in Section 3.3.

- Character-by-character normalization
  - Character code normalization[*8]
  - Normalize half-width characters / full-width characters
  - Delete double spaces
  - Delete spaces at beginning / end of sentences
- Inconsistent bilingual text processing
  - Mismatched proper noun substitution[*9]

[*7] In this paper, “consistency of the meaning of parallel texts” means that there is no conflict of meaning between the English translation of the Japanese source text and the relevant English sentence, and that there is also no conflict of meaning between the Japanese translation of the English source text and the relevant Japanese sentence (meaning that there is consensus when translating forward and backward between Japanese and English).

[*8] Japanese CG Reports are provided with Unicode text data (XBRL), but English is only available in PDF format and character codes may not match between Japanese and English. In addition, CJK Compatibility Ideographs (U+F900 to U+FAFF) may be included in Japanese sentences, so these were normalized to the representative characters of CJK unified ideographs. (Example: U+F9BE (料) → U+6599 (料))

[*9] In cases where the Japanese sentence uses pronouns or omits such grammatical elements as the subject but the English translation uses proper nouns instead, the proper noun has been replaced with a pronoun or generic term. For example, if the subject of

  - Remove unmatched symbols[*10]
  - Delete notes accompanying translation
- Removal of parallel texts, etc. that include words or phrases considered inappropriate for machine translation by NMT model
  - Delete URLs
  - Delete strings with only personal names or numbers
- Delete duplicate parallel texts

In this paper, the parallel corpus created through these processes is defined as the "Parallel Corpus of CG Reports". The number of sentence pairs included in the Parallel Corpus of CG Reports was approximately 43,000. The distribution of the string length of Japanese sentences and the number of English words is shown in Fig. 3.

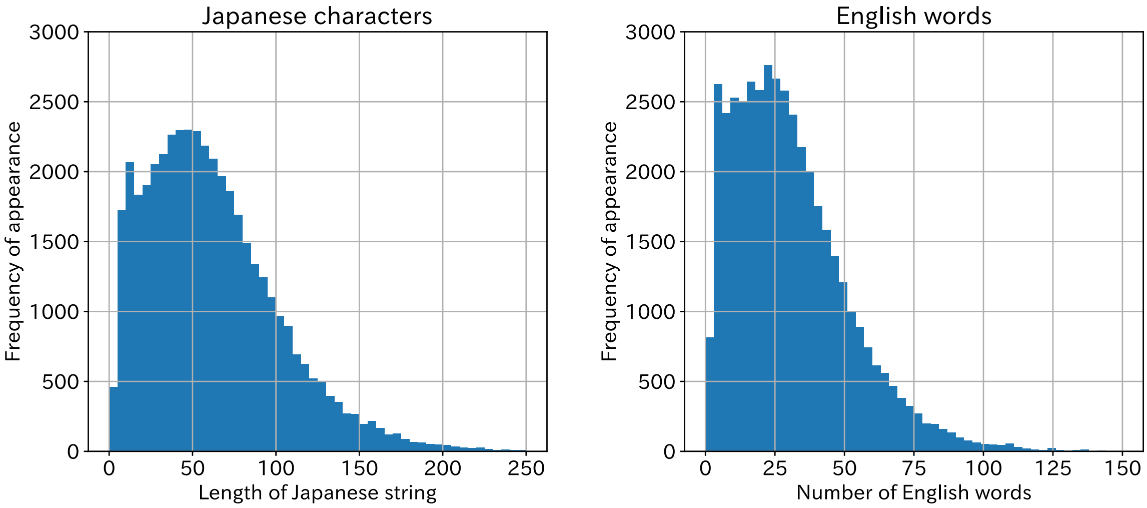


Fig. 3 Distribution of Japanese character string length (left) and distribution of number of English words (right)

5 Experiment

5.1 Setting

In this study, experiments were performed using 1) AutoML Translation[12][*11] provided by Google Cloud Platform™ and 2) another cloud-based machine translation service ("Service A") that provide creation of custom translation models.

AutoML Translation is a domain adaptation service for Google NMT[13] ("GNMT") that uses an input

---

the Japanese sentence is "当社" but the English translation uses the actual company name, the company name has been replaced with the term "The Company".

[*10] Refers to tasks that involve deleting symbols in parallel texts where either only the Japanese or only the English contains such symbols as parentheses or asterisks

[*11] As of January 15, 2019, beta version.

parallel corpus to create a custom translation model tailored to said parallel corpus*[12]. The custom translation models created by AutoML Translation and Service A are machine translation models in which the sentences from the source language are input one at a time and corresponding translations are output ("1-to-1 translation model"). The output is generated as isolated translated sentences, regardless of the context of input source text.

Two parallel corpora were used in this experiment. The first, "Full Corpus", is the Parallel Corpus of CG Reports built for this study. The second, "Abridged Corpus", is a parallel corpus that was generated after deleting certain sentence pairs from the Full Corpus: those with a Japanese character string length equal to or greater than the threshold (100 characters) and sentence pairs with the number of English words equal to or greater than the threshold (50 words). These thresholds were determined based on the value considered to be the inflection point in the Fig. 3, as a value that could not ensure a sufficient number of translated sentence pairs considering the character string length and the number of words. In addition, since bilingual texts with 50 or more English words are automatically deleted in Service A, verification was performed using only the Full Corpus. Table 7 shows the number of parallel translations used in this training.

Three experiments were performed for each bilingual corpus. Evaluations based on BLEU scoring were carried out for each experiment, and then the average values were calculated.

Table. 7 Breakdown of experimental data

Service names	Parallel Corpus	Train	Dev	Test	Total
AutoML Translation	Full Corpus	34,527	4,316	4,315	43,158
	Abridged Corpus	27,641	3,455	3,455	34,551
Service A	Full Corpus	38,446	2,129	2,129	42,704

5.2 Results and discussion

Table. 8 BLEU score in each experiment setting (averages)

Service names	Input Parallel Corpus	Custom version	Base line	Amount of increase
AutoML Translation	Full Corpus	25.572	19.631	+5.941
	Abridged Corpus	26.022	19.781	+6.241
Service A	Full Corpus	28.277	21.620	+6.657

We made a baseline for each regular service and compared the BLEU score with a custom translation model that applied domain adaptation to each service. The results are shown in Table 8.

With the custom translation model, BLEU scores increased by about 6 points in all cases. In addition, the increase in BLEU scores was larger for custom translation models of the Abridged Corpus.

Table 9 shows the results of inputting examples of the Japanese sentences used for evaluation into the custom translation model created from the Abridged Corpus and the normal GNMT. Although the input

---

*[12] Specific algorithms for domain adaptation performed by AutoML Translation and Service A are not disclosed.

Table. 9 Examples of output of machine translation model

1	Input	株式会社ICJの提供する議決権電子行使プラットフォームに参加しております。
	Reference translation	The Company is participating in the platform for the electronic exercise of voting rights operated by ICJ, Inc.
	Base line	We participate in the platform for exercising voting rights electronics provided by ICJ Co., Ltd.
	Custom ver.	The Company participates in the electronic voting platform provided by ICJ, Inc.
2	Input	2016年度において、当社の会計監査人は新日本有限責任監査法人であります。
	Reference translation	The Accounting Auditor of the Company for the year ended March 31, 2017 is Ernst & Young ShinNihon LLC.
	Base line	In fiscal 2016, our accounting auditor is New Japan limited liability audit corporation.
	Custom ver.	In fiscal 2016, the Company’s Accounting Auditor was Ernst & Young ShinNihon LLC.
3	Input	当社グループは、企業価値の最大化を目指し、コーポレート・ガバナンスの徹底を最重要課題の一つと位置付け様々な施策を講じています。
	Reference translation	The Group is aiming to maximize corporate value, and has been implementing various measures as rigorous corporate governance is our highest priority.
	Base line	The Group aims to maximize corporate value and positions thorough corporate governance as one of the most important tasks and takes various measures.
	Custom ver.	The Group considers thoroughness of corporate governance as one of the most important issues with the aim of maximizing corporate value, and takes various measures.
4	Input	さらに同氏は取締役会議長として、当社グループの経営の基本方針等について、取締役会としての決議に向け議案審議を主導いたしました。
	Reference translation	Furthermore, as the Chairman of the Board of Directors, she led the Board of Directors to make decisions on proposals, including a proposal for basic management policy of the Group.
	Base line	In addition, as Chairman of the Board of Directors, he led the deliberation on the agenda for resolutions as the Board of Directors regarding the Group’s basic management policies and others.
	Custom ver.	In addition, as Chairman of the Board of Directors, he led the deliberation of the Board of Directors on the basic policy on the management of the Group.
5	Input	外部識者による講演会の開催、社内WEBサイトでの情報発信、座談会実施等による啓発活動
	Reference translation	Holding awareness campaigns through round-table discussions, publishing of information on the Company’s internal website, and hosting lectures by visiting experts.
	Base line	Held lectures by outside experts, disseminate information on internal website and raise awareness through implementation of round-table discussion etc.
	Custom ver.	Held lectures by outside experts, disseminate information on the internal website, and conduct awareness-raising activities such as holding round-table talks.
6	Input	最近1年間において、(イ)または当社の業務執行者に該当していた者
	Reference translation	A person who falls into (ii) or was an executor of business for the Company in the last year.
	Base line	In the past year, (a) or those who corresponded to the Company’s business executor.
	Custom ver.	A person who has fallen under either of the above (a) or an executive of the Company’s business for the past year.
7	Input	当社グループの経営理念は、「いつも、人から。」です。
	Reference translation	The Group’s management philosophy is “Putting People First.”
	Base line	The management philosophy of the Group is "always from people."
	Custom ver.	The Group’s management philosophy is "always from a person."

sentences in Examples 1 and 2 contain technical terms and proper nouns, it can be confirmed that the custom translation model was able to output correct translations. Also, looking at the translation results in Example 3, the baseline translates “最重要課題” into “the most important tasks”, while the custom translation model outputs “the most important issues”. The latter is a more appropriate English translation given the context.

On the other hand, the other examples in Table 9 are cases where the custom translation model created in this experiment shows no improvement in translation quality. Example 4 includes “同氏”, (See Section 3.2), which is “she” in the reference translation but was output as “he” under both the baseline and custom translation models. Example 5 includes “等” (See Section 3.4), and although the translation result of the custom translation model is better than that of the baseline when looking at the sentence as a whole, neither output modifies the same words as the reference translation. Example 6 includes “(イ)” (See Section 3.5), which is “(ii)” in the reference translation but was output as “(a)" under both the baseline and custom translation models. Example 7 includes coined words and poetic expressions (See Section 3.6), and the nuances of the input sentences are lost in the translation results of the baseline and custom translation models. Context and background knowledge of each expression is required to correctly translate such source text. As such, it is thought that this issue would be difficult to resolve with domain adaptation in the 1-to-1 translation model.

6 Conclusion

In this study, we built a parallel corpus using existing CG Reports, and then evaluated results of a custom translation model based on said corpus to examine the possibility of using machine translation for CG Reports. Our experiments confirmed that BLEU scores increased when English translations of technical terms, proper nouns, and other expressions were modified. The parallel corpus built in this study is relatively small, which means that further expansion of and improvements to the quality of the corpus can result in further improvements to translation quality. However, for translating sentences that require background knowledge or context, it is considered necessary to consider machine translation models that take these matters into consideration, pre-processing and post-processing, etc. We intend to continue expanding on this study.

Based on the above, we propose the following areas for future studies.

- Challenges on machine translation
  - Studying machine translation models that incorporate context
  - Studying pre-processing and post-processing of symbols for ordered lists
  - Studying of machine translation models that adapt to text type (For instance: those that can intuitively assess tags from XML data to determine text categories such as short/long sentences and proper nouns, which can then be used to determine which machine translation model, including rule-based or NMT, should be used, depending on the text category)
- Broader challenges
  - Improving quality of and expanding parallel corpora in the securities and IR fields
  - Studying glossaries and style guides in the securities and IR fields
  - Studying the guidelines for alignment, pre-processing, and data cleansing when building parallel corpora in specific domains
  - Systematically surveying methods for writing Japanese that is suitable for NMT models

Acknowledgements

We wish to express our deep appreciation for the valuable feedback we received from Yusuke Oda of Google Japan LLC and Prof. Kumiko Tanaka-Ishii at Research Center for Advanced Science and Technol-

ogy, The University of Tokyo.

References

[1] 土井惟成, 近藤真史, 山藤敦史. コーポレート・ガバナンス報告書における機械翻訳の検討. 言語処理学会第 25 回年次大会 (NLP2019), pp. 926–929, 3 2019.

[2] 株式会社東京証券取引所, 株式会社名古屋証券取引所, 証券会員制法人福岡証券取引所, 証券会員制法人札幌証券取引所. ２０１７年度株式分布状況調査結果の概要.

[3] 株式会社 東京証券取引所情報サービス部. 投資部門別売買状況. `https://www.jpx.co.jp/markets/statistics-equities/investor-type/00-02.html`, 2017. (参照 2019-01-15).

[4] 投資家フォーラム. 投資家フォーラム-第１・２回会合-報告書. `https://investorforum.jp/`. (参照 2019-01-15).

[5] 株式会社東京証券取引所. 東証上場会社 コーポレート・ガバナンス白書, 2017.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[7] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39. Association for Computational Linguistics, 2017.

[8] 株式会社みらい翻訳. 職場英語力を TOEIC 900 点相当に引き上げる日英双方向機械翻訳サービスをリリース. `https://miraitranslate.com/uploads/2017/12/befdfe2e9eca64235a2042cd9f50a3db.pdf`, 12 2017. (参照 2019-01-15).

[9] Translation API - Dynamic Translation | Translation API. `https://cloud.google.com/translate/`. (参照 2019-01-15).

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation, 2002.

[11] 藤田篤, 山田優, 影浦峡. 産業翻訳に役立つ自然言語処理技術についての議論の足場. 言語処理学会第 25 回年次大会 (NLP2019), pp. 914–917, 3 2019.

[12] AutoML Translation | Google Cloud. `https://cloud.google.com/translate/automl/docs/`. (参照 2019-01-15).

[13] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, Vol. abs/1609.08144, , 2016.