Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3719–3726
Marseille, 11–16 May 2020

TDDC: Timely Disclosure Documents Corpus

Nobushige Doi*, Yusuke Oda, Toshiaki Nakazawa†**
* Japan Exchange Group
2-1, Nihombashi-kabuto-cho, Chuo-ku, Tokyo, 103-8224, Japan
n-doi@jpx.co.jp

** Google Research
3-21-3, Shibuya, Shibuya-ku, Tokyo, 150-0002, Japan
oday@google.com

†The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan
nakazawa@logos.t.u-tokyo.ac.jp

Abstract

In this paper, we describe the details of the Timely Disclosure Documents Corpus (TDDC). TDDC was manually organized by aligning the sentences from past Japanese and English timely disclosure documents in PDF format published by companies listed on the Tokyo Stock Exchange. TDDC consists of approximately 1.4 million parallel sentences in Japanese and English. TDDC was used as the official dataset for the 6th Workshop on Asian Translation to encourage the advancement of machine translation.

Keywords: Parallel corpus, Machine translation, Asian language, Stock exchange, Investor Relations

1. Introduction

The Tokyo Stock Exchange (TSE) is one of the largest capital markets in the world, with over 3,600 companies listed as of the end of 2018. TSE–listed companies are required to disclose material information to the public in a timely manner. This information is written in timely disclosure documents and includes financial statements, corporate actions, and corporate governance policies. Moreover, the documents are essential for investment decisions and are disclosed on the TSE websites[1].

From the surveys by TSE, the proportion of overseas corporation ownership of Japanese listed company shares (based on market capitalization) has risen almost consistently since 1990 to currently around 30%[2]. As of 2018, the proportion of share trading value coming from transactions by overseas investors is approximately 59%[3]. Although tens of thousands of original Japanese documents are disclosed every year (i.e., over 79,000 documents in 2018), the availability of English disclosure documents is limited. Thus, there is a strong demand for machine translation on both listed companies and global investors since Japanese to English translation needs to be done in a timely manner.

However, it is difficult for TSE–listed companies to translate all their documents owing to the volume of information, limited resources of translators, and time constraints. The amount of text in timely disclosure documents tends to be large. In 2018, the total number of pages in all timely disclosure documents (over 79,000 documents disclosed by over 3,600 companies) exceeded 710,000; which means on average, a TSE–listed company is required to disclose over 197 pages each year. For TSE–listed companies, to translate all these pages would take huge amounts of time and money. Moreover, there are not enough translators available for all timely disclosure documents because the demand for translation clusters in peak season. For example, in 2018, approximately 48% of TSE–listed companies disclosed annual earnings reports (over 38,000 pages) in the same week, from May 9th to 15th. Furthermore, most investors require TSE–listed companies to disclose both Japanese and English documents simultaneously. Consequently, it is not easy to meet the demand for the English translation of timely disclosure documents using manual translation only. The machine translation could therefore be a solution to these problems.

For current machine translation systems aimed at a specific field, a parallel corpus adapted to that field is regarded as an essential resource. There are already Japanese–English parallel corpora for training machine translation systems in fields such as patents (Utiyama and Isahara, 2007) and scientific papers (Nakazawa et al., 2016). However, thus far, there is no large-scale Japanese–English parallel corpus specifically aimed at the Investor Relations field.

The Timely Disclosure Documents Corpus (TDDC) consists of approximately 1.4 million Japanese–English sentence pairs that have been extracted from past timely disclosure documents and other documents. Timely disclosure documents contain important figures (e.g., sales, profits, and dates) and proper nouns (e.g., names of people, places, companies, businesses, and products). This information is essential for investors; thus, mistranslations need

[1] https://www.jpx.co.jp/english/listing/disclosure/index.html

[2] https://www.jpx.co.jp/english/markets/statistics-equities/examination/01.html

[3] https://www.jpx.co.jp/english/markets/statistics-equities/investor-type/00-02.html

to be avoided, and overall translation quality needs to be improved.

TDDC was prepared by Japan Exchange Group (JPX), which is an operator of securities exchanges including TSE. It was provided for research at the 6th Workshop on Asian Translation (WAT)[4]. During the 6th WAT, TDDC was free and available online only for research on natural language processing such as machine translation[5]. In this paper, we introduce details of TDDC and briefly explain the findings from the 6th WAT perspective.

2. Timely Disclosure Task

Timely disclosure task, which is one of the subtasks for the 6th WAT, aims to improve the Japanese to English translation of sentences extracted from timely disclosure documents to avoid mistranslations that would confuse investors. As terms on which investors focus, we define two groups: important figures and proper nouns. These terms cannot be accurately translated using typical Neural Machine Translation (NMT) systems because the NMT systems restrict the vocabulary size and consider rare words (e.g., names and numbers) as out-of-vocabulary words (Luong et al., 2015). The current NMT systems introduce the subword tokenization, which transfers rare words to the sequence of its constituent characters (Sennrich et al., 2016). However, the subword tokenization solves the problem only if a rare word can be translated as constitutive words. Thus, even using the subword tokenization, the NMT systems often are often unable to translate neither numbers with many digits nor constitutive proper nouns. The following sections will explain the summary of the timely disclosure documents and their details.

2.1. Timely Disclosure Documents

Timely disclosure documents are disclosed in PDF format on the TSE websites. There are mainly three categories of timely disclosure documents: performance results, corporate governance reports, and documents that describe material facts pertaining to business or other matters of listed companies. The material facts gets defined in the Japanese law (Article 166 of the Financial Instruments and Exchange Act).

TSE–listed companies are usually required to disclose their performance results as "決算短信" (*Kessan tanshin*, Earnings reports) four times a year and corporate governance reports, which describe their status of corporate governance at least once a year. These periodic or annual documents tend to contain many common words and sentences for each company because companies tend to write new documents by referring to their past documents. Meanwhile documents that describe material facts are disclosed as required, not periodically.

2.2. Important Figures

We define important figures as numbers that have financial meaning (e.g., not page numbers and item numbers) such as amounts and dates.

Table 1: Examples of amounts

Japanese	English
224,812 円	224,812 yen
10 億円	1 billion yen
1,283,929 千円	1,283,929 thousand yen
105 億 37 百万円	10,537 million yen

Table 2: Examples of dates

Japanese	English	Unbalanced Translation
平成 30 年 6 月 26 日	June 26, 2018	-
2018 年度第 1 四半期	the first quarter of fiscal 2018	-
2019 年 3 月期	FY March 2019	-
2018 年 5 月期	FY May 2018	Q1 FY2018
本年 4 月 1 日	April 1 of this year	April 1, 2018
当第 3 四半期連結累計期間	consolidated cumulative third quarter of this fiscal year	3Q FY03/2018

Examples of amounts include sales, revenue, and numbers of sales. The numbers in English increase by thousands; however, the Japanese numerals group numbers by 10,000. There are various patterns of units of amounts in Japanese timely disclosure documents, particularly. The following patterns "1 億 5 千万円" ($1 \times 10^8 + 5 \times 10^7$ yen), "1 億 5000 万円" ($1 \times 10^8 + 5,000 \times 10^4$ yen), "1 億 50 百万円' ($1 \times 10^8 + 50 \times 10^6$ yen), and "150 百万円" (150×10^6 yen) indicate "150 million yen". Examples of amounts are shown in Table 1.

Examples of dates include document issue dates, accounting period, and fiscal year. The imperial era name (or Japanese era name), such as "平成" (*Heisei*) and "令和" (*Reiwa*) are widely used in Japanese documents for counting years, instead of Anno Domini (AD). There are cases where some companies omit information on dates in Japanese but provide them in English. For example, although "前期" (prior period) was written in a Japanese document, an English document described it as "fiscal year 2017". In Japanese timely disclosure documents, there are many variable prefixes for dates such as "本" (this), "当" (this), "次" (next), and "前" (previous, last). Examples of dates are presented in Table 2.

For performing the timely disclosure task, context-based accurate translation is not necessary (e.g., omissions of words in Japanese sentences and abbreviated numbers). However, these unbalanced sentences cause poor corpus quality.

2.3. Proper Nouns

We define proper nouns as rare words that are found only in documents from one company such as names of people, company names, names of places, and product names. This definition limits its original meaning and excludes technical terms such as accounting and legal terms. These terms can be found not only in the documents from one company but also in the documents from other companies. For investors, it is important to clearly grasp the subjects and objects in sentences, which should not be misplaced or translated in other proper nouns.

Moreover, in timely disclosure documents, there are pairs of sentences for which the information in Japanese and En-

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html

[5] http://lotus.kuee.kyoto-u.ac.jp/WAT/Timely_Disclosure_Documents_Corpus/

glish are not equivalent owing to the differences in proper nouns between Japanese and English sentences: omission of subjects or objects and addition of proper nouns. There are cases where some companies omit the subject and object in Japanese but supplement proper nouns in English. In other cases, in Japanese timely disclosure documents, some companies frequently use pronouns; however, in English, those pronouns are replaced with proper nouns. Similar to dates described in Section 2.2, for this task, it is not necessary to achieve accurate translation based on the context (e.g., translation of Japanese pronoun to proper nouns). Table 3 shows examples of subject or object omissions and proper nouns additions.

3. Construction of the Corpus

The construction process of TDDC consists of the following steps: (1) gathering source documents, (2) aligning documents and sentences, and (3) preprocessing sentences.

3.1. Source Documents

The source documents of TDDC are documents that satisfied the following conditions:

- disclosed from January 1, 2016 to June 30, 2018
- written by TSE–listed companies or Real Estate Investment Trusts (REITs)
- disclosed in both Japanese and English, and
- not encrypted or rasterized (i.e., sentences can be extracted).

The source documents include documents that were disclosed daily (not timely), such as Corporate Governance Reports, which are also essential for investors.

3.2. Aligning Documents and Sentences

The pairs of documents and pairs of sentences were manually aligned through crowdsourcing. The main procedure is as follows: (a) A worker picks an English document from the documents which no one has picked yet. (b) The worker finds the corresponding Japanese document. (c) The worker extracts English sentences from top to bottom of the English documents and sees each of the similar Japanese sentences. (d) A Checker examines the alignment results. Thus the sentences of one document are aligned by only one worker, and there is no automatic alignment generated before the manual alignment.

Sentences that were difficult to align were excluded (e.g., translations with notes in English), and not all sentences in each timely disclosure document were included in TDDC.

Although checkers carefully examined the alignment results, there may have been the following errors: character corruption, alignment errors, and not extracted characters at the beginning or end of sentences.

3.3. Preprocessing

The aligned sentences underwent five preprocessing steps to remove noises (i.e., character corruption, control characters, and extra spaces).

Replacing characters Most TSE–listed companies made timely disclosure documents using Windows OS computers because this is the recommended environment of TSE systems. Although files created in Japanese Windows OS are mainly encoded with CP932 (Code Page 932), timely disclosure documents frequently contained characters that are not defined in CP932. Therefore, some specific characters are replaced with other characters described in CP932.

Unicode normalization There are two ways of expressing alphanumeric characters in Japanese sentences on computers: full-width and half-width. To normalize characters (including the abovementioned characters), the sentences are normalized with NFKC (Normalization Form Compatibility Composition)[6], with the following exceptions: Numbers enclosed within a circle ("①"–"⑳": U+2460–U+2473), Two dot leaders ("‥": U+2025), and Horizontal ellipsis ("…": U+2026). The words in these exceptions are often written in timely disclosure documents, and each of them and their normalized characters will have distinct meanings. For example, numbers enclosed within a circle will be normalized with NFKC into integers, and both these numbers will be used in the original document as distinct item numbers.

Deleting control characters Sentences extracted from PDF documents sometimes contain control characters. Therefore, control characters are removed such as the characters whose Unicode Character Categories are "Cc," "Cf," "Cn," or "Co."

Deleting extra spaces Extra spaces in the sentences are deleted such as spaces at the beginning and end of sentences, and more than one space.

Deleting non–Japanese–English pairs To make TDDC contain Japanese–English pairs, non–Japanese–English pairs are deleted such as an English sentences that contains Japanese characters and Japanese sentences that does not contain Japanese characters.

4. Dataset

TDDC is partitioned into the training (Train), development (Dev), development–test (DevTest), and test (Test) data. The sets of source documents used as training, development, development–test, and test data are independent of each other. Furthermore, each data set of the development, development–test, and test is further split into two sets of data. Sentences that end with a Japanese period ("。": U+3002) are classified as `Texts`, which have various sentences, and others are classified as `Items`, which contain many duplicates and similar expressions.

4.1. Data Format

TDDC consists of Japanese–English sentence pairs, document hashes, and sentence hashes. A document hash is a hash of the `Document ID` (`DID`), which is a unique identifier of the source document. A sentence hash is a hash of the `DID` and `Sentence ID` (`SID`), which is a unique identifier of the sentence in each source document. Pairs of

[6] `http://www.unicode.org/reports/tr15/`, as of Nov. 2019

Table 3: Examples of proper nouns (proper nouns and corresponding words are shown in bold)

Japanese	English
取締役兼代表執行役グループ CEO **清田 瞭**	**Akira Kiyota**, Director & Representative Executive Officer, Group CEO
清算関連収益は、**株式会社日本証券クリアリング機構**が行う金融商品債務引受業に関する清算手数料等から構成されます。	Clearing services revenue comprises clearing fees related to the assumption of obligations of financial instrument transactions carried out by **Japan Securities Clearing Corporation**.
また、育児・介護支援制度の充実を図り、仕事との両立ができるよう環境整備に取り組んでいます。	**JPX** has also enriched its childcare and caregiving leave systems to create an environment that allows employees to balance work and family commitments.
発行済株式数に占める**当社**保有株式の比率	Shareholding ratio of **JPX**
当社の企業理念及び社会的使命に共感していただけるとともに、金融行政に関する豊富な経験と高い見識を**当社**の経営に反映することができるため、社外取締役に選任しています。	**Mr. Tsuda** has been appointed as an outside director due to his capacity to identify with the Company's corporate philosophy and social mission as well as his considerable experience and insight on financial policy and systems which can be expected to be reflected into the management of **JPX**.

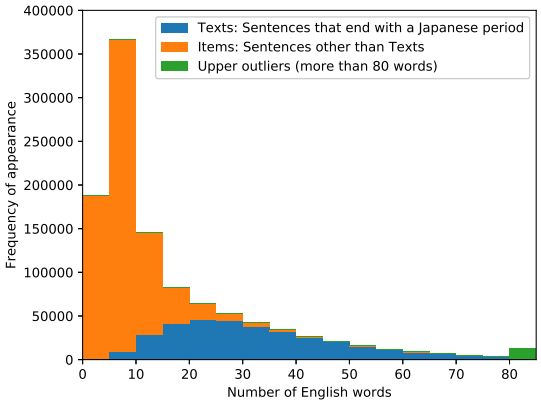

Figure 1: Distribution of the number of English words in the sentences of Train–2016–to–2017

DID and SID guarantee that these Japanese–English sentence pairs are independently extracted from the original documents. Sample of rows of TDDC that contain DID, SID, Japanese sentences, and English sentences are described in Table 4.

4.2. Data Splitting

The training data is split into two sets of data from different periods. The first data set was created based on documents disclosed from January 1, 2016 to December 31, 2017 (Train-2016-to-2017), and the second data set was based on documents from January 1, 2018 to June 30, 2018 (Train-2018).

Development, development–test, and test data sets were extracted from timely disclosure documents disclosed from January 1, 2018 to June 30, 2018, excluding documents that were used to create the training data. The documents for the period were randomly selected, and the sentences were extracted from each randomly selected, discrete document set; thus, the extracted sources were not biased. Moreover, annotators modified English sentences to guarantee that the pairs of development, development–test, and test data were semantically correct.

5. Analysis of TDDC

Table 5 shows the statistics of TDDC. To understand the difficulty of the timely disclosure task, we conducted an analysis of TDDC.

A challenging issue for the general NMT systems is long sentence processing (Bahdanau et al., 2014). Approximately 23% of sentences in Train-2016-to-2017 are long (i.e., the number of words is over 50). Figure 1 shows the distribution of the number of English words in the sentences of Train–2016–to–2017.

Table 5 shows that there are many duplicate sentences in each dataset. Table 6 shows the number of duplicate sentences in Train-2016-to-2017. It gets confirmed that there are more duplicate Japanese sentences than English sentences, and there are more duplicate sentences in Items than in Texts.

Table 7 gives the distribution of source documents in Train-2016-to-2017. Similar to Section 3.1, TDDC consists of sentences from timely disclosure documents that are disclosed both in Japanese and English; thus the publishing companies are biased. The number of the publishing TSE–listed companies in Train-2016-to-2017 is 590, although 3,602 companies are listed on TSE as of the end of 2017.

6. Timely Disclosure Task at the 6th Workshop on Asian Translation

TDDC was provided for the 6th WAT, which is an open evaluation campaign that focuses on Asian languages. The participants of the timely disclosure task can submit the results of Texts and/or Items. During the 6th WAT, the translation performance of the results underwent automatic and human evaluations (Nakazawa et al., 2019). As automatic evaluation, the following three metrics were used: BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015). For human evaluation, two types of evaluations were used: pairwise crowdsourcing evaluation (Nakazawa et al., 2016) and Japan Patent Office (JPO) adequacy evaluation. In addition to these official evaluations during the 6th WAT, to focus on the timely disclosure task, we particularly evaluated the results of JPO adequacy evaluation.

6.1. JPO Adequacy Evaluation

The JPO adequacy evaluation was performed by translation experts with a quality evaluation criterion for translated patent documents that was decided by the JPO. Table 8 shows the JPO adequacy criterion from 5 to 1. The evaluation was performed subjectively. "Important information" represents the technical factors and their relationships. The degree of importance of each element was also evaluated. The percentages in each grade are rough indications for the transmission degree of the source sentence meanings. The detailed criterion is described in the JPO document (in

Table 4: Sample rows of Texts (sentences that end with a Japanese period) and Items (other sentences) in TDDC

DID	SID	Japanese	English
/ID3l...	n3jL...	当社は、2017 年 10 月 30 日に開示しました 2018 年 3 月期 (2017 年 4 月 1 日～2018 年 3 月 31 日) の通期連結業績予想及び期末の 1 株当たり配当予想について、下記のとおり修正することとしましたので、お知らせいたします。	We hereby announce that the consolidated earnings forecast and year-end dividend forecast for the fiscal year ending March 31, 2018 released on October 30, 2017 have been revised as follows.
hnK...	NsGw...	これにより、2018 年 3 月期の期末の 1 株当たり配当金は、普通配当 33 円に加え、記念配当 10 円を合わせた 43 円となります。	As a result, the year-end dividend per share for the fiscal year ended March 31, 2018 will be ¥ 43 (ordinary dividend of ¥ 33 plus commemorative dividend of ¥ 10).
98lr...	PG52...	投資活動によるキャッシュ・フローは、無形資産の取得による支出 105 億 37 百万円等により、261 億 64 百万円の支出となりました。	There was cash outflow of ¥ 26,164 million from investment activities due mainly to ¥ 10,537 million in purchase of intangible assets.
TGHF...	XDD6...	SGX が保有する自己株式 (515,063 株) を含む。	Including the shares held by SGX as treasury stock (515,063 shares).

DID	SID	Japanese	English
TGHF...	fhWJ...	株式会社日本取引所グループ	Japan Exchange Group, Inc.
/ID3...	dlms...	業績予想及び配当予想の修正に関するお知らせ	Notice of Revision to Earnings Forecast and Dividend Forecast
hnKj...	Juml...	剰余金の配当に関するお知らせ	Notice of Dividend from Surplus
TGHF...	EdLv...	発行済株式数に占める当社保有株式の比率	Shareholding ratio of JPX

Table 5: Number of sentences in TDDC (the values given in parentheses indicate the number of unique pairs)

Disclosure Period	Train	Dev		DevTest		Test	
		Texts	Items	Texts	Items	Texts	Items
2016-01-01 to	1,089,346	-	-	-	-	-	-
2017-12-31	(614,817)	-	-	-	-	-	-
2018-01-01 to	314,649	1,153	2,845	1,114	2,900	1,153	2,129
2018-06-30	(218,495)	(1,148)	(2,650)	(1,111)	(2,671)	(1,135)	(1,763)

Table 6: Duplicate sentences in Train–2016–to–2017

		Number	Percentage
Texts	Duplicated pairs	75,606	21.74%
	Duplicated Japanese	107,029	30.77%
	Duplicated English	84,803	24.38%
	Whole pairs	347,788	-
Items	Duplicated pairs	398,923	53.81%
	Duplicated Japanese	504,207	67.99%
	Duplicated English	429,398	57.90%
	Whole pairs	741,558	-

Table 7: Distributions of source documents and publishing companies in Train–2016–to–2017

	Documents	Publishers
From TSE–listed companies	9,950	590
Performance Results	4,311	493
Corporate Governance Reports	347	143
Material Facts	4,688	321
Others	604	100
From REITs	2,713	50
The whole of the source	12,663	640

Japanese)[7].

6.2. Evaluation of Important Figures and Proper Nouns

The number of test sentences for the JPO adequacy evaluation was 200. A total of 200 test sentences were randomly and carefully selected from the test dataset to focus on the translation evaluation of important figures and proper nouns defined in Section 2. We rated each of the sentences in terms of important figures and proper nouns as

[7] http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

Table 8: The JPO adequacy criterion

Score	Meaning
5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%–)
3	More than half of important information is transmitted correctly. (50%–)
2	Some of important information is transmitted correctly. (20%–)
1	Almost all important information is NOT transmitted correctly. (–20%)

Table 9: Results of evaluation of important figures and proper nouns (%)

		ntt	NICT-2	sarah	geoduck
Proper Nouns	Texts (N=50)	68.0	68.0	72.0	76.0
	Items (N=50)	74.0	68.0	74.0	-
Important Figures	Texts (N=92)	93.5	93.5	92.4	68.5
	Items (N=83)	97.6	96.4	94.0	-

0 or 1, where 1 means that the words (important figures or proper nouns) are correctly translated.

6.3. Participants

During the 6th WAT, 4 teams participated in the Japanese–English timely disclosure documents task (Nakazawa et al., 2019). Morishita et al. as `ntt` (Morishita et al., 2019), Imamura and Sumita as `NICT-2` (Imamura and Sumita, 2019), and Susanto et al. as `sarah` (Susanto et al., 2019) used NMT without other resources. Eriguchi et al. as `geoduck` (Eriguchi et al., 2019) used translation memory and NMT with 1 million Japanese–English Wikipedia parallel corpus provided by Asai et al. (Asai et al., 2018) as an additional training resource.

6.4. Evaluation Results

Figures 2 and 3 denote the official results of the timely disclosure task (Nakazawa et al., 2019). Table 9 shows the evaluation results of the sentences containing important figures or proper nouns. Table 10 shows the sample results of

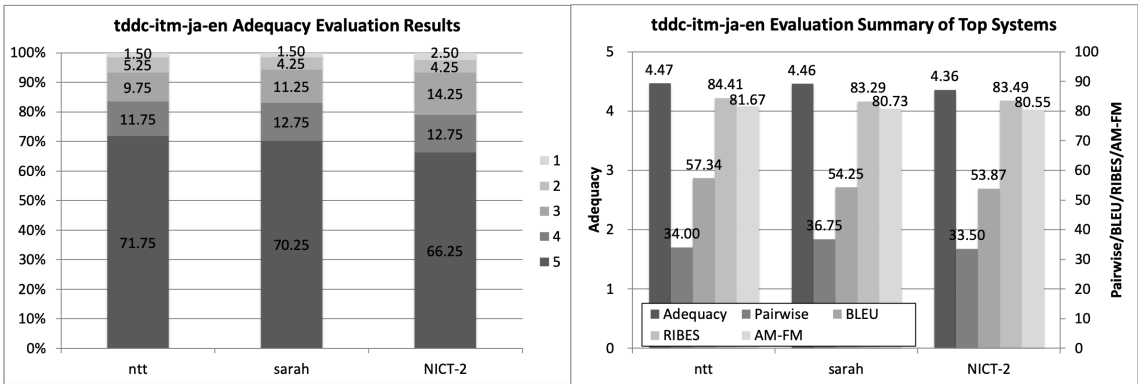

Figure 2: Official evaluation results of Items (tddc-itm-ja-en)

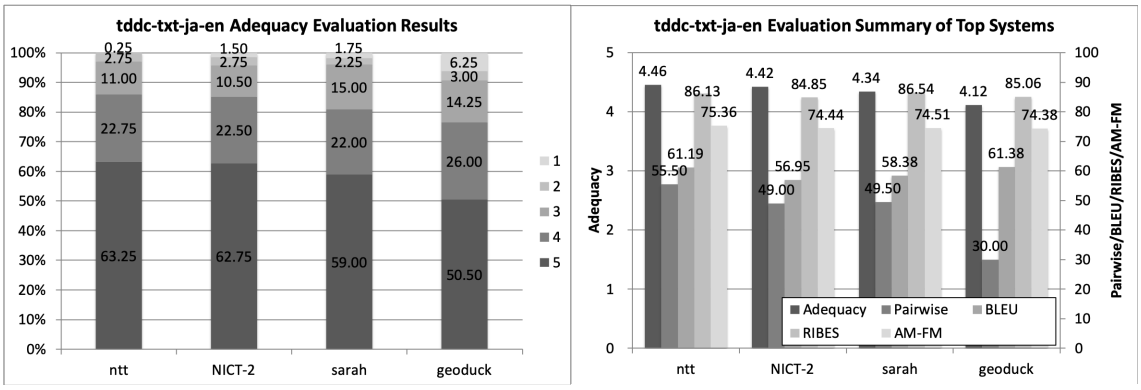

Figure 3: Official evaluation results of Texts (tddc-txt-ja-en)

participants.

In the results, all systems achieved approximately 4 points according to the JPO adequacy evaluation scores for both `Items` and `Texts`, and all evaluators rated over 70% of all pairs as 4 or 5. We examined these results and determined that most of these high-rated pairs consist of typical terms and sentences from timely disclosure documents, including long sentences.

The sentences of Examples 1, 2, and 3 in Table 10 include important figures or proper nouns; in addition, the source sentences in Example 3 are long, and they were correctly translated. Moreover, the source sentences in Example 1 and 3 contain complex numbers with Japanese numerals; however, they were correctly translated by the systems. It is assumed that these sentences are typical in timely disclosure documents, and there are sufficient sentences for training models.

Despite these scores, however, Table 9 shows that there are mistranslations of important figures and proper nouns.

6.4.1. Mistranslations

We determined that there were four patterns of mistranslations in these results: uncommon expressions, appearance of unrelated proper nouns, incorrect modifiers or determiners, and sentences that contained interpreted numbers. Table 10 shows various error types, which are analyzed below:

Uncommon sentences and words used in timely disclosure documents tend to be mistranslated. The structure of the source sentences in Example 4 seemed to be uncommon in timely disclosure documents, and some sentences were scored low. In Example 4, some figures in the sample results were mistranslated and modified with irrelevant date. Moreover, in Example 5, names of people were rare in TDDC, and they were mistranslated. The abovementioned information implies that the translations of uncommon sentences are considerably affected by sentences in the training data that are similar but have different meaning.

Some systems tended to translate sentences without subjects into sentences with incorrect subjects. As mentioned in Sections 2.2, Japanese sentences frequently omit subjects and objects that would normally be included in English. However, source sentences in Examples 6 and 7 that included “当社” (The Company) were sometimes translated to unrelated company names. Similarly, in Example 8, despite the lack of the subject, the translated sentence contained a specific personal name. To achieve accurate translation by machine translation systems, awareness of the context is required, otherwise unnatural or passive sentences are output. It is assumed that there were the same or similar sentences in the training data, and the subjects in English sentences contained proper nouns.

There are some incorrect modifiers or determiners in these results. As mentioned Sections 2.1, in Japanese timely disclosure documents, there are many variable prefixes for dates. Some systems translated sentences containing these words with an incorrect year. For example, the source sen-

tence in Example 9, “当第 3 四半期連結会計期間末” (the end of third quarter of this fiscal year) was translated as “the end of the third quarter of FY 2015” or “the end of the third quarter of FY 2016”.

In TDDC, there are sentences that contain interpreted numbers. For example, translation of the year from the imperial era name to AD requires machine translation systems to understand the conversion rules. Some systems appear to be able to translate these dates, but not every time, as shown in Example 7. As another example, a Japanese sentence in a document states “X の伸長率は 99.4%となりました。” (Growth rate of X was 99.4%.); however, in an English document, it is stated that “X declined 0.6%.”. The interpretation of numbers in Japanese–to–English translations is frequently seen in timely disclosure documents. In Example 10, these pairs in TDDC, which required the interpretation of numbers, were mistranslated.

In conclusion, the following causes for these mistranslations are considered:

- It is difficult to translate long sentences and proper nouns that TDDC does not contain.
- Some source sentences are unclear owing to the lack of subjects and/or objects; thus, there is no suitable English translation without a context.
- TDDC contains not semantically balanced pairs, and the systems may be considerably affected by either source pair sentences.

6.4.2. Evaluation Subjectivity

We determined two patterns of improper JPO adequacy scores owing to the subjectivity of evaluators: false positive and false negative.

Some translations results include unrelated proper nouns and incorrect modifiers or determiners; however, some evaluators did not seem to consider those as deductions, such as Examples 6 to 9. We call these higher than expected scores as false positive.

Some translations seem to be appropriate for sentences in TDDC that contain paraphrasing expressions; however, evaluators gave them low scores probably because they were not literal translations. For example, some systems correctly translated sentences that included interpreted numbers, but evaluators did not appreciate this and deducted points. In Example 10, one evaluator gave a score of 4 to the sentence that was identical to the reference sentence; however, the other evaluator gave a score of 1. We call these lower than expected scores as false negative.

These false positive and false negative tendencies imply that it is necessary to create another evaluation criterion, which correctly evaluates the correctness of transmitted information to investors.

7. Conclusion

In this study, we introduced TDDC, which is a large-size parallel corpus of timely disclosure documents. The purpose of TDDC is to contribute to the improvement in machine translation of sentences in these documents. However, we predict that TDDC could be diverted for use in benchmarks for Named–Entity Recognition because the sentences in TDDC have particular useful characteristics with respect to proper nouns and figures. The results of the 6th WAT suggest that most sentences that are typical in TDDC and do not depend on context are translated correctly. However, there are mistranslations in sentences that contain words that are not present in TDDC or whose meaning changes depending on the context. Further consideration is needed to improve these translations, such as an expansion of language resources, context-aware machine translation systems, and suitable evaluation criteria for the timely disclosure task. TSE–listed companies disclose many timely disclosure documents every year. We should consider using the rich source of information in these documents to expand language resources in the Investor Relations field.

8. Bibliographical References

Asai, A., Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. (2018). Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banchs, R. E., D'Haro, L. F., and Li, H. (2015). Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):472–482, March.

Eriguchi, A., Rarrick, S., and Matsushita, H. (2019). Combining translation memory with neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 123–130, Hong Kong, China, November. Association for Computational Linguistics.

Imamura, K. and Sumita, E. (2019). Long warm-up and self-training: Training strategies of NICT-2 NMT system at WAT-2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 141–146, Hong Kong, China, November. Association for Computational Linguistics.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July. Association for Computational Linguistics.

Morishita, M., Suzuki, J., and Nagata, M. (2019). NTT neural machine translation systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 99–105, Hong Kong, China, November. Association for Computational Linguistics.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language*

Table 10: Example of translation results (values given in parentheses indicate the average values of the JPO adequacy evaluations obtained using two evaluators)

Ex.		Sentences
1	Source	取締役の基本報酬の総額は、年額 **10 億円** (うち、社外取締役は **5,000 万円以内**) としております。
	Reference	The total amount of basic remuneration for Directors is set at **1.0 billion yen** per year (of which, **up to 50 million yen** for Outside Directors).
	Result (5.0)	The total amount of basic remuneration for Directors is **￥1.0 billion** per year (including **￥50 million or less** for Outside Directors).
	Result (5.0)	The total amount of base compensation for Directors is **1 billion yen** per year (including outside directors of **50 million yen** per year).
2	Source	**第一三共**が **2016 年 6 月**に申請
	Reference	Submitted by **Daiichi Sankyo** in **June 2016**
	Result (5.0)	Submitted by **Daiichi Sankyo** in **June 2016**
	Result (5.0)	Submitted by **Daiichi Sankyo** in **June 2016**
3	Source	これは主に、増加要因として税金等調整前当期純利益が **12 億 31 百万円**、その他の資産の増減によるキャッシュ・フローが **6 億 79 百万円**、仕入債務の増減によるキャッシュ・フローが **6 億 7 百万円**それぞれ増加した一方、減少要因として退職給付に係る負債の増減によるキャッシュ・フローが **13 億 42 百万円**減少したことによるものであります。
	Reference	As positive factors, profit before income taxes rose **1,231 million yen**, and the increase in other assets and the increase in notes and accounts payable - trade totaled **679 million yen** and **607 million yen**, respectively. As a negative factor, cash flow from increase and decrease in net defined benefit liability fell **1,342 million yen**.
	Result (5.0)	This was primarily the result of increases in income before income taxes of **1,231 million yen**, cash flows from changes in other assets of **679 million yen** and cash flows from an increase or decrease in notes and accounts payable-trade of **607 million yen**, while there was a decrease in net defined benefit liability of **1,342 million yen**.
	Result (5.0)	This was mainly due to increases of **1,231 million yen** in income before income taxes, **679 million yen** in cash flows from changes in other assets and **607 million yen** in cash flows from changes in notes and accounts payable-trade, while there was a decrease of **1,342 million yen** in net defined benefit liability.
4	Source	この結果、保有契約の年換算保険料*1 は **10,796 百万円**となるとともに、保有契約件数は **255,618 件**となりました。死亡保険の保有契約高は **2,028,255 百万円**となりました。
	Reference	Accordingly, annualized premium*1 of policies-in-force was **10,796 million yen**. The number of policies-in-force resulted in a total of **255,618**, and sum insured of policies-in-force stands at **2,028,255 million yen**.
	Result (4.0)	Accordingly, annualized premium*1 of policies-in-force was **10,796 million yen**. The number of policies-in-force as of **the end of July 2017** resulted in a total of **255,618**, and sum insured of policies-in-force stands at **2,028,255 million yen**.
	Result (3.0)	Accordingly, annualized premium*1 of policies-in-force was **10,210 million yen**. The number of policies-in-force as of **the end of May 2017** resulted in a total of **242,379**, and sum insured of policies-in-force stands at **1,976,419 million yen**.
5	Source	※ 1 **菊田徹也**および**瓜生宗大**の取締役就任は、当局認可を前提とします。
	Reference	(*1)…The appointment of **Tetsuya Kikuta** and **Munehiro Uryu** are based on the premise of an approval from the authority.
	Result (4.0)	*1 Based on the approval of the authorities, **Kengo Sakurada** and **Kengo Oshiro** will assume the office of Director.
	Result (4.0)	*1 The appointment of **Tetsuya Kikita** and **Mr. Uryu Uryu** is subject to the approval of the regulatory authorities.
6	Source	**当社**は、株主の皆様に対する利益還元を経営の重要課題の一つと位置づけております。
	Reference	**The Company** recognizes the return of profit to its shareholders as a key management priority.
	Result (4.5)	**Kyowa Hakko Kirin** regards the return of profits to its shareholders as one of its key management priorities.
	Result (4.5)	**FANCL** considers the distribution of profit to shareholders to be an important management issue.
7	Source	**当社**は、**平成 30 年 3 月 31 日**を基準日とする剰余金の配当 (期末配当) について、以下の内容で本日、取締役会決定致しましたのでお知らせします。
	Reference	**The Company** has announced that at a meeting held today, the Board of Directors passed a resolution to pay dividends from surplus (year-end dividend) with a record date of **March 31, 2018**.
	Result (4.5)	**Teijin Limited ("the Company")** has announced that at a meeting held today, the Board of Directors passed a resolution to pay dividends from surplus (year-end dividend) with a record date of **March 31, 2018**.
	Result (4.0)	**Teijin Limited ("The Company")** has announced that at a meeting held today, the Board of Directors passed a resolution to pay dividends from surplus (year-end dividend) with a record date of **March 31, 2017**.
8	Source	お客さま視点でのマーケティングに長けたグローバル企業の経営のトップとして、豊かな経験と経営に関する高い見識を有しております。
	Reference	**He** has extensive experience and deep insight into management as a top management of a global company which is proficient at marketing from the customers' perspectives.
	Result (4.5)	**Mr. Ito** has extensive experience and deep insight into management as a top management of a global company which is proficient at marketing from the customers' perspectives.
	Result (4.5)	**Mr. Ito** has extensive experience and deep insight into management as a top management of a global company which is proficient at marketing from the customers' perspectives.
9	Source	**当第 3 四半期連結会計期間末**における純資産合計は 229,856 百万円となり、**前連結会計年度末**に比べ 97,228 百万円減少いたしました。
	Reference	Total net assets at **the end of third quarter of this fiscal year** were 229,856 million yen, a decrease of 97,228 million yen versus **the end of the previous fiscal year**.
	Result (4.5)	Total net assets at the end of the third quarter of **FY 2016** were 229,856 million yen, a decrease of 97,228 million yen versus the end of **FY 2015**.
	Result (4.5)	Total net assets at the end of the third quarter of **FY 2017** were 229,856 million yen, a decrease of 97,228 million yen from the end of **FY 2016**.
10	Source	うち、主力分野の BtoB 事業の**伸長率は 101.7%**、LOHACO の**伸長率は 99.4%**となりました。
	Reference	Of the total, non-consolidated net sales of mainstay B-to-B business **grew 1.7%** on a year-on-year basis and those of LOHACO **declined 0.6%**.
	Result (2.5)	Of the total, non-consolidated net sales of mainstay B-to-B business **grew 1.7%** on a year-on-year basis and those of LOHACO **declined 0.6%**.
	Result (1.0)	Of the total, non-consolidated net sales of mainstay B-to-B business **grew 5.3%** on a year-on-year basis and those of LOHACO **declined 7.1%**.

Resources and Evaluation (LREC'16), pages 2204–2208, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., Bojar, O., and Kurohashi, S. (2019). Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China, November. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Susanto, R. H., Htun, O., and Tan, L. (2019). Sarah's participation in WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 152–158, Hong Kong, China, November. Association for Computational Linguistics.

Utiyama, M. and Isahara, H. (2007). A japanese-english patent parallel corpus. In *In proceedings of the Machine Translation Summit XI*.