上場会社開示資料の英訳文における文分割に関する考察

A Consideration for Sentence Splitting in English Translations of Timely Disclosure Documents

株式会社日本取引所グループ　総合企画部フィンテック推進室

土井　惟成

2015 年株式会社日本取引所グループに入社。東京証券取引所 IT 開発部などを経て、2018 年より現職。

n-doi@jpx.co.jp

1 はじめに

東京証券取引所（以下、東証）は 3,700 社を超える企業（2021 年 9 月時点）が上場している世界最大の証券市場の一つである。東証上場会社は、投資家の投資判断に影響を与えうる情報を、東証が運営する Web システム（以下、TDnet）を通じて開示することが義務付けられている。本稿では、上場会社が TDnet を通じて開示する書類を「開示資料」と呼ぶ。開示資料の規模は膨大であり、2020 年における日本語の開示資料は約 10.6 万文書、総ページ数は約 86 万ページに及んでいる。そのため、開示資料の読者にとっては、膨大なデータから投資判断上有用な情報を取得するために、機械翻訳や自動要約といった技術を活用し、情報処理に係る負担を抑えたいというニーズがあるものと推察する。

しかしながら、開示資料中の本文等のテキストは、文構造が複雑なものが多く、機械翻訳や自動要約をはじめとする自然言語処理技術の精度が落ちやすい。開示資料からなる日英対訳コーパスである TDDC [1] を用いた分析では、100 文字以上の和文は全体の約 21%、50 単語以上の英文は全体の約 17% を占めている。このヒストグラムを【図 1】に示す。

このような長文の処理精度を高める方法として、対象の長文を予め可読性の高い文に言い換えることが挙げられる。先行研究[2]では、開示資料中のテキストにおける、人手による言い換え手法を検討した結果、短文化が特に有効だったことを示した。従って、このような可読性を高める前処理が機械的に実現されれば、開示資料におけ

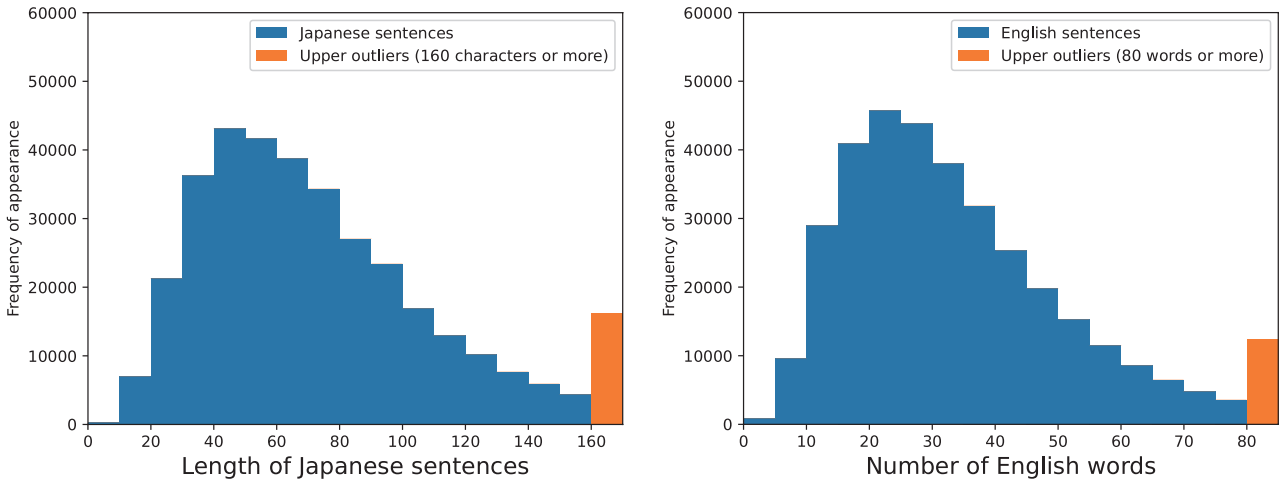

図 1　2017 年 -2018 年の開示資料から構築した日英対訳コーパス（TDDC）のうち、和文が句点（。）で終わる対訳（約 34.7 万文対）における和文文字列長（左図）と英文単語数（右図）のヒストグラム

る自然言語処理技術の活用が期待されると考える。

そこで、開示資料に適した文分割（Sentence Splitting）の手法に係る検討を目的として、開示資料の英訳文における文分割に関する考察を述べる。本稿における文分割とは、同じ意味を保つように１文を複数の短い文に変換することを指す。以下では、先行研究[2]で使用した、開示資料の一つであるコーポレート・ガバナンスに関する報告書（以下、CG 報告書）からなる日英対訳コーパスにおいて、和文１文に対して英文が複数の文に分割されている対訳を抽出し、文分割の傾向を分析した。

2 関連研究

先行研究[2]では、開示資料の機械翻訳の翻訳品質向上の手段として、産業日本語への言い換えの有効性の検証を行った。具体的には、開示資料の一つである CG 報告書を対象に、産業日本語への言い換え作業をクラウドワーカー等に依頼のうえ、機械翻訳モデルへの入出力結果を比較した。この時、言い換えルールとして「特許ライティングマニュアル第２版」[3]を利用した。検証の結果、【表 1】に示すように、長文の文分割が特に有効であったことを確認した。一方で、文分割の作業結果の中には、分割後の主語や目的語に誤った語句を補足したことで、英訳に誤訳が生じた事例が散見された。

文分割は、テキスト平易化（Text Simplification）の構成技術の一つと見なすことができる。テキスト平易化とは、難しい文を同じ意味の平易な文に変換することであり、大きく、語彙の平易化と文構造の平易化の２段階で構成される。その中で文分割は、文構造の平易化の一つとして位置づけられている。日本語の自動文分割の手法としては、文を適切な箇所で区切り、主語、文末表現、接続詞を補完するという手法が提案されている[4、5]。また、最近では、seq2seq モデルを用いた文分割の手法が提案されている[6]。日本語の開示資料についても、言い換えコーパスを構築することで、同様のモデルの実現が期待できるものと推察する。また、このようなモデルを機械翻訳の前処理として用いることで、機械翻訳の精度向上に寄与することが期待される。

網川[7]は、特許明細書を対象に、英訳文が分割されている対訳を抽出し、英訳文を再翻訳することで言い換えコーパスを得るとともに、これを利用した言い換え方法を提案している。開示資料の文分割についても当該手法が準用できるものと推察する。

表１　長文の文分割による翻訳品質の改善例（太字箇所の言い換えにより、下線箇所の訳抜けが改善）

原文	当社は、任意の指名・報酬委員会などの独立した諮問委員会を設置していませんが、取締役候補の選任や取締役の報酬については、**取締役会の決議に先立ち**、独立社外取締役及び親会社に対し説明を行い、適切な助言を得ております。
参照訳	Although NTT DATA has not set up an independent advisory committee such as a voluntary nomination committee or remuneration committee, **in advance of the resolution by the Board of Directors**, we provide independent directors and the parent companies with explanations about the nomination of candidates for directors and remuneration for directors and receive appropriate advice from them.
原文 →機械翻訳	The Company has not established an independent advisory committee such as a voluntary nomination and remuneration committee. We give explanations and obtain appropriate advice.
言い換え文	当社は、任意の指名・報酬委員会などの独立した諮問委員会を設置していません。**ですが、当社は、**取締役候補の選任や取締役の報酬については、取締役会の決議に先立ち、独立社外取締役及び親会社に対し説明を行い、適切な助言を得ています。
言い換え文 →機械翻訳	The Company has not established an independent advisory committee such as a voluntary nomination and compensation committee. **However, prior to the resolution of the Board of Directors**, the Company explains the selection of director candidates and director remuneration to independent outside directors and the parent company to obtain appropriate advice.

3 英訳文分割コーパスの抽出

本稿では、先行研究 [2] で用いた日英対訳コーパスを元に、1 文に対して英文が複数の文に分割されている対訳で構成される対訳コーパス（以下、英訳文分割コーパス）を作成し、これについて分析を行う。英訳文分割コーパスの元となる日英対訳コーパスは、2019 年 7 月までに日本語及び英語で開示された CG 報告書からパラグラフ単位でランダムに抽出した、全 591 文対の日英対訳コーパスである。その中で、和文が 100 文字以上の文対（322 文対）の内、英文が複数文に分割されている 156 文対を、本稿での分析対象となる英訳文分割コーパスとした。

4 文分割のパターン

英訳文分割コーパスにおける文分割の傾向を分析するために、本稿では、特許ライティングマニュアルにおける、言い換えルールカテゴリ 1 の「短文にする」を参考に、【表 2】のとおり、以下の 4 種類の文分割のパターンを定義する。

（1）文節での分割
（2）文節間距離の圧縮
（3）箇条書きの分離
（4）括弧書きの分離

本稿の分析においては、英訳文分割コーパスの対訳を比較し、英訳文での分割手法が（1）-（4）のどのパターンに最も近いか、人手によって分類した。

5 分析及び考察

英訳文分割コーパス（156 文対）における文分割のパターンを分類した結果を【表 3】に示す。また、各文分割パターンの例を、英文の再翻訳と併せて【表 4】に示す。以下では、それぞれの文分割パターンでの傾向等についてそれぞれ述べる。

まず、全体の約 90.4% における分割手法が「（1）文節での分割」に分類された。この手法は、和文を適切な箇所で区切り、主語、文末表現、接続詞を補完するという手法である。英訳文分割コーパス中の和文には、【表 4】の例 1-1 及び例 1-2 のように、「～だが」や「～おり」といった接続節の直後で文が分割されている傾向があった。そのため、和文から分割箇所を予測してから英文を見ることで、区切られている文節や補完すべき語句を特定することは容易だった。このことから、このような対訳を参考とすることで、文分割の言い換えコーパスを、手作業で構築することは、比較的低コストで実現できるものと推察する。

一方で、【表 4】の例 1-2 の再翻訳文のように、同じ主語（当社）が連続することで文が単調となり、流暢さが損なわれているような印象を受ける事例が散見された。そのため、「（1）文節での分割」による文分割においては、こういった流暢さの確保が課題の一つとして挙げられるものと推察する。

表 3　文分割のパターンの分布（全対訳数＝ 156）

	名称	対訳数	割合
（1）	文節での分割	141	90.4%
（2）	文節間距離の圧縮	14	9.0%
（3）	箇条書きの分離	1	0.6%
（4）	括弧書きの分離	0	0.0%

表 2　文分割のパターンの定義

名　称	手法の定義	特許ライティングマニュアルにおける対応ルール
（1）文節での分割	文を適切な箇所で区切り、主語、文末表現、接続詞を補完	1-2: 複数の主語や述語を含むときは、文を分ける
（2）文節間距離の圧縮	係り受けが離れている文節同士を近づけ、間に含まれる修飾語等を以降の文へ分離	1-1: 説明語句が長いときは、短く分ける
（3）箇条書きの分離	文中に箇条書きを含む場合、箇条書きを以降の文へ分離	1-3: 箇条書きは文を分ける
（4）括弧書きの分離	文中に長い括弧書き含む場合、括弧書きを以降の文へ分離	1-4: 文中の長いかっこ書きは分ける

表４　文分割の例（太字箇所は和文と再翻訳文の差分を表す）

例	項目	文
1-1	和文	当社は、本報告書の提出時点において、女性の役員を選任しておらず、本原則が実施できておりません**が、**取締役会および監査役会の人員構成において、ジェンダー面も含む多様性が求められていることの重要性を認識しており、役員候補者について、女性を含む多様性を確保できるように今後検討してまいります。
	英文	As of the date of submission of this report, no female director has been elected by the Company and therefore this Principle has not been implemented. However, we recognize the importance of having diversity, including gender diversity, in the composition of the Board of Directors and the Audit & Supervisory Board, and will consider director candidates so as to ensure diversity to include female candidates.
	再翻訳文	当社は、本報告書の提出時点において、女性の役員を選任しておらず、本原則が実施できておりません。**しかしながら、当社は、**取締役会および監査役会の人員構成において、ジェンダー面も含む多様性が求められていることの重要性を認識しており、役員候補者について、女性を含む多様性を確保できるように今後検討してまいります。
1-2	和文	当社では経営幹部の指名・報酬については経営会議、取締役の指名・報酬については、複数名かつ取締役会の３分の１以上を占める独立社外取締役が出席する取締役会において十分に審議して**おり、**現時点では任意の指名･報酬委員会を設置しておりません。
	英文	The Company sufficiently deliberates nomination and remuneration of senior management at the Management Committee and nomination and remuneration of Directors at the Board of Directors meeting attended by multiple Independent Outside Directors accounting for one-third or more of the total number of Directors. The Company has no voluntary nomination or remuneration committee.
	再翻訳文	当社では経営幹部の指名・報酬については経営会議、取締役の指名・報酬については、複数名かつ取締役会の３分の１以上を占める独立社外取締役が出席する取締役会において十分に審議して**います。当社は、**現時点では任意の指名・報酬委員会を設置しておりません。
2-1	和文	取締役会は、事業の執行状況を適切に理解し、機動的、且つ迅速な意思決定と執行状況の監督をできるよう、業務上の経験・知識・専門性を有する社内取締役と、ステークホルダーや社会の求める視点を踏まえ、問題提起を行うことができる複数の社外取締役により構成することを基本方針としております。
	英文	Our basic policy is to have a board of directors consisting of inside directors and a multiple number of outside directors. Inside directors have operational experience, knowledge, and expertise to properly understand the Company’ s business operations, provide agile and timely decision-making, and oversee the execution of our business. Outside directors provide insight from stakeholders and society, raising issues to be addressed by the Company.
	再翻訳文	取締役会は、**社内取締役と複数の社外取締役により構成することを基本方針としております。社内取締役は、**事業の執行状況を適切に理解し、機動的、且つ迅速な意思決定と執行状況の監督をできるよう、業務上の経験･知識･専門性を有**します。社外取締役は、**ステークホルダーや社会の求める視点を踏まえ、問題提起を行**います。**
2-2	和文	当社におけるコーポレート･ガバナンスの充実･強化については、株主のみなさまをはじめ、お客さま、取引先、債権者、地域社会等の様々な利害関係者の利益の最大化、ならびに当社の持続的な成長と中長期的な企業価値の向上を目的として、重要な戦略の実行にあたり、透明性、公正性および迅速性を確保したうえで、前例や慣習にとらわれない果断な意思決定を行うための機能と、業務執行に対する監督機能の強化という点を重要課題として認識し、各種施策に取り組んでおります。
	英文	With regard to enhancing and strengthening the Company's corporate governance, the Company is implementing various measures with the recognition that it is vital to strengthen the function to make decisive decisions unshackled by precedents or customs as well as to strengthen the supervisory function for business execution, by ensuring transparency, fairness and speed when carrying out key strategies. The aim is to maximize the interest of various stakeholders, including our shareholders as well as our customers, business partners, creditors and local communities, and achieve sustained growth as well as enhance the medium- to long-term corporate value of the Company.
	再翻訳文	当社におけるコーポレート・ガバナンスの充実・強化については、**重要な戦略の実行にあたり、透明性、公正性および迅速性を確保したうえで、前例や慣習にとらわれない果断な意思決定を行うための機能と、業務執行に対する監督機能の強化という点を重要課題として認識し、各種施策に取り組んでおります。これにより、**株主のみなさまをはじめ、お客さま、取引先、債権者、地域社会等の様々な利害関係者の利益の最大化、ならびに当社の持続的な成長と中長期的な企業価値の向上を**目指します。**

「(2) 文節間距離の圧縮」は、全体の約 9.0% に留まった。この手法による文分割は、【表 4】の例 2-1 及び例 2-2 のように、「(1) 文節での分割」よりも全体的な文構造が大きく変わる傾向が見られた。従って、この手法による文分割の件数が少ない理由として、文分割の労力が大きいことや、著者が日英での文構造の一致を強く求めるといった事情が挙げられると推察する。

一方で、【表 4】の例 2-1 のように、「(2) 文節間距離の圧縮」による文分割は、文構造がより明瞭となることで、文の正確さと流暢さを大きく向上させる印象を受けた。そのため、文分割による可読性の向上を追求する場合、このような言い換え事例を集積し、この分割手法について検討を深めることが手段の一つとして考えられる。

「(3) 箇条書きの分離」は 1 件であり、「(4) 括弧書きの分離」は 0 件だった。この理由として、このような文中の箇条書きや括弧書きは、特許文書には頻出ではあるものの、開示資料ではあまり見受けられないといった傾向があるのではないかと推測する。

6 おわりに

本稿では、開示資料の英訳文における文分割の傾向について分析を実施した。この分析の結果、開示資料の英訳時における文分割の傾向として、文節での分割が大部分を占めることが判明した。また、このような英文を再翻訳することで、開示資料中のテキストの言い換えコーパスが構築できるという示唆を得た。今後は、本稿での分析を足がかりに、開示資料中のテキストの平易化等に向けて引き続き調査を進めたいと考える。

参考文献

[1] Nobushige Doi, Yusuke Oda, and Toshiaki Nakazawa. TDDC: Timely Disclosure Documents Corpus. In Proceedings of The 12th Language Resources and Evaluation Conference（LREC 2020）, pp. 3719-3726, Marseille, France, 5 2020.

[2] 土井惟成，大西恒彰，百石弘澄，高頭俊，山藤敦史 . 上場企業開示資料の機械翻訳におけるプリエディットの検討 . 言語処理学会第 26 回年次大会（NLP2020），pp. 525-929, 3 2020.

[3] 一般財団法人日本特許情報機構特許情報研究所 . 特許ライティングマニュアル「産業日本語」, 第 2 版, 3 2019.

[4] 金淵培，江原暉将 . 日英機械翻訳のための日本語長文自動短文分割と主語の補完 . 情報処理学会論文誌 . Vol. 35, No. 6. pp.1018-1028.（1994）

[5] 美野秀弥，田中英輝 . やさしい日本語ニュースのための自動文分割 . 言語処理学会第 19 回年次大会（NLP2013）, pp. 264-267, 3 2013.

[6] Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. Split and rephrase. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 606-616, Copenhagen, Denmark, 9 2017.

[7] 綱川隆司 . ニューラル機械翻訳における長文翻訳のための文分割による言い換え方法の検討 . Japio YEAR BOOK 2020. pp. 288-291, 11 2020.