文分割による読みやすさへの影響に関する考察

土井 惟成　　大西 恒彰　　命苫 昭平　　嶋根 正輝　　高頭 俊
株式会社 日本取引所グループ
{n-doi, n-onishi, s-meitoma, m-shimane, s-takato}@jpx.co.jp

概要

文構造が複雑な長文を読みやすくする方法として、長文を複数の短文に分割する文分割 (Sentence Splitting) という手法が挙げられる。本研究では、上場会社開示資料中の文を対象として、人手による長文の文分割が、読みやすさに関する指標にどのような影響を及ぼすか考察する。実験では、評価指標として、サプライザルの総和や係り受け木の深さを利用した。この結果、文分割による長文の文分割によって、係り受け木の深さが平均 36%程度低減することを示した。

1　はじめに

東京証券取引所 (以下、東証) は、3,800 社を超える企業 (2022 年 1 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。公正で透明な株価形成の確保の目的から、東証上場会社には、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) を通じて適時適切に公表する義務が課されている。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と推察する。しかしながら、開示資料には長文をはじめとする文構造が複雑な文が多く、機械的な処理の精度が低くなりやすい [1]。

文構造が複雑な長文を読みやすくする方法として、長文を複数の短文に分割する文分割 (Sentence Splitting) が挙げられる。開示資料中の文の機械翻訳による英訳精度の向上を目的とした、人手による前処理手法を検討した研究では、文分割による短文化が特に有効だったことを示した [2, 3]。一方で、文分割によって文にどのような変化が生じているかについては、定量的には明らかになっていない。

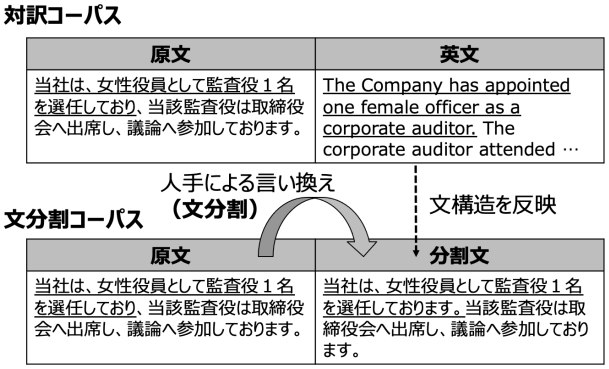

図 1　文分割コーパスの作成手順の概要

本研究では、開示資料中の文を対象として、人手による長文の文分割が、読みやすさに関する指標にどのような変化を及ぼすかについて考察する。具体的には、開示資料の一種である CG 報告書から作成した日英対訳コーパス（以下、CG 報告書対訳コーパス）[2] をもとに、英文における文分割を踏まえつつ、人手にて原文に文分割を施し、小規模なコーパスを作成した。本稿では、原文に文分割を施すことで得られた複数の文を総称して分割文、原文と分割文で構成されるコーパスを文分割コーパスと呼ぶ。文分割コーパスの作成の流れを図 1 に示す。そして、小規模な文分割コーパスを対象に読みやすさに関する指標を算出し、それぞれを比較することで、文分割に係る示唆を得ることを目指す。

2　関連研究

2.1　文分割 (Sentence Splitting)

先行研究 [4] では、開示資料の日英対訳コーパスにおける、1 文に対して英文が複数の文に分割されている対訳に着目し、その傾向を分析した。具体的には、CG 報告書対訳コーパスから、英文が複数文

表 1　文分割のパターン

#	パターン名称	文分割の手法
1.	文節での分割	文を適切な箇所で分割し、主語、文末表現、接続詞を補完
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割
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割
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割

に分割されている対訳を抽出し、和文と英文における文構造の変化を、特許ライティングマニュアル [5] における、言い換えルールカテゴリ 1「短文にする」の言い換えルールに当てはめた。この文分割のパターンの内容を表 1 に示す。本研究では、当該研究の手順を参考にして、文分割コーパスの作成手順を策定した。

日本語における機械的な文分割の手法として、係り受け解析の結果を元に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 [6]。この手法は、表 1 における「1. 文節での分割」に相当すると考えられる。また、近年では、英文における機械的な文分割の手法として、seq2seq モデルを用いた手法が提案されている [7]。文分割のパターンの網羅や大規模な文分割コーパスの作成によって、日本語の開示資料についても高精度な機械的な文分割が期待できると推察する。

2.2　読みやすさの指標

文の読みやすさの指標は、自然言語処理システムの出力結果の品質評価をはじめとして、幅広い用途で利用されている [8]。その中でも、内容の理解に必要なリテラシー (学校教育年数) を表したものが、読みやすさの評価指標として広く利用されている。英語におけるこのような評価指標を推定する手法として、Gunning Fog Index[9] や Flesch-Kincaid [10] がある。これらの手法では、文の平均単語数と単語の複雑性をもとに、米国の学年レベルに対応するスコアを算出する。また、Schwarm らは、言語モデルの複雑さ (perplexity) や係り受け木の深さ等の素性の有用性を報告している [11]。日本語を対象とした同様の評価手法としては、日本語の教科書コーパスに基づいて導出された、1 文字ごとの集計値 (unigram) を用いた Sato らの評価式 [12] や、形態素等を用いた Shibasaki らの評価式 [13] がある。

また、読みやすさの評価指標として、文の読み時間を用いた研究も行われている [14]。近年、日本語の書き言葉の読み時間に係る研究では、新聞記事データの文に読み時間を付与したコーパス (BCCWJ-EyeTrack) が利用されている [15]。文の読み時間と関連する指標として、サプライザル理論に基づくサプライザルが指摘されている [16, 17, 18]。Kuribayashi は、BCCWJ-EyeTrack をもとにサプライザルのモデリングを行い、人間の文処理の計算モデルと言語モデルの関連について報告している [19]。

3　データセット: 文分割コーパス

本研究では、先行研究 [2] を踏まえ、CG 報告書対訳コーパスをもとにした、人手による文分割コーパスの作成手順を策定した。CG 報告書対訳コーパスは、2019 年 7 月までに日本語及び英語で開示された CG 報告書から抽出した、全 591 文対の日英対訳コーパスであり、和文が 100 文字以上の文で構成されている。文分割コーパスの作成手順は次のとおりである。

1. 改行を含む対訳や、原文と英文で文意が均衡していない対訳を削除する。
2. 原文 1 文に対して、英文が複数の文に分割されている対訳文を抽出する。
3. 原文と英文の文構造を比較し、表 1 を元に文分割のパターンを判断する。
4. 英文の文構造を踏まえつつ、文の流暢さと正確さを損なわないように、原文に文分割を施す。

本研究では、この手順に従って、全 146 文対の文分割コーパスを作成した。文分割コーパスの統計情報を表 3、原文における文字列長の出現頻度の分布を図 2 に示す。表 3 のとおり、文分割コーパスに含まれる分割文のパターンは「1. 文節での分割」と「2. 文節間距離の圧縮」のみであった。これらの文分割の例を表 2 にそれぞれ示す。以下では、手順 4. の文分割で見られた傾向について、文分割のパターンごとに述べる。

「1. 文節での分割」では、原文と英文で語順が同一であることが多かった。そのため、英文から分割すべき箇所が判明すれば、当該箇所で文を区切り、主語、文末表現、接続詞を補完することで文分割は完了した。したがって、人手による文分割の作業としては、比較的容易であった。なお、本手順においては、原文から文意が乖離しない限りにおいて、補足する語句は英文に合わせた。ただし、補足する主

表 2　文分割の例 (分割文における強調箇所は原文との差分を表す)

#	区分	文
1	原文	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。
	分割文	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、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。
2	原文	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。
	分割文	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。

表 3　文分割コーパスの統計情報 (文対数以外は平均値)

項目	パターン 1. 文節での分割	パターン 2. 文節間距離の圧縮	全体
文対数	128	18	146
文字列長 (原文)	148.7	160.7	150.1
編集距離	8.4	75.7	16.7

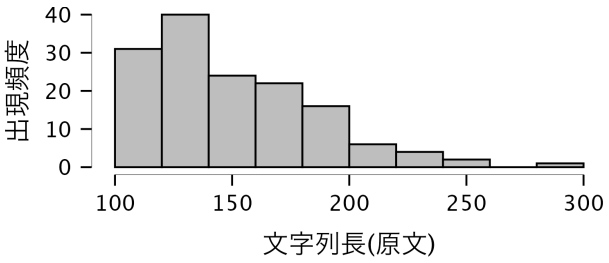

図 2　文分割コーパスにおける原文の文字列長の分布

語が長過ぎる場合（例：会社名）は、文の流暢さを優先して、後半の文の主語は省略するといった対応を行った。

「2. 文節間距離の圧縮」では、原文と英文で語順が異なるため、文全体の構造を踏まえつつ、節の係り受けや分割する箇所を検討する必要がある。その結果として、表 3 のとおり「1. 文節での分割」よりも編集距離が大きくなり、文分割の作業としては比較的労力が大きいと言える。

4　実験

4.1　実験設定

本実験では、文分割コーパスをもとに、原文と分割文に対して読みやすさに関する指標を算出し、それぞれの指標の変化を評価する。読みやすさに関する指標として、文字列長、トークン数、係り受け木の深さ、サプライザルを採択した。文字列長とトークン数は、文の複雑さを簡易的に測る指標として広く使われており、特にトークン数は文の読みやすさと有意に相関することが報告されている [20, 21]。係り受け木の深さは、文構造の複雑さを表す指標であり、読みやすさとの関連が報告されている [20, 11]。サプライザルは、トークンごとの読み時間に関連すると考えられている指標である [16]。本実験では、文分割によってサプライザルの総和は低減するという仮説を立て、これを指標として採択した。なお、読みやすさの指標として、語彙に関連する指標 (例: 単語の難しさ等) が広く使われるが、文分割では語彙の平易化は行わないことから、これらの利用は見送った。

トークン数の算出に当たっては、後述するサプライザルと同様に、MeCab[22] と UniDic を用いて入力文のトークン化を行った。サプライザルは、Kuribayashi のコード[1)] [19] を用いてトークン単位のサプライザルを算出し、その総和を指標として用いた。なお、分割文の文字列長、トークン数、サプライザルを算出する時は、文境界で分割せず、分割後の複数の文を 1 つの入力文とした。トークン単位のサプライザルの算出結果の例を図 3 に示す。

係り受け木の深さの算出には、GiNZA[2)] による係り受け解析を行った。分割文については、文境界で分割し、各短文に対して係り受け解析を行い、係り受け木の深さの最大値を指標として用いた。

4.2　実験結果及び考察

原文と分割文における各指標の評価結果を表 4 に示す。まず、全体的な傾向として、文分割によって文字列長とトークン数が増加傾向にあった。この理由として、文分割では一文ごとの文字列長やトーク

1)　https://github.com/kuribayashi4/surprisal_reading_time_en_ja
2)　https://megagonlabs.github.io/ginza/

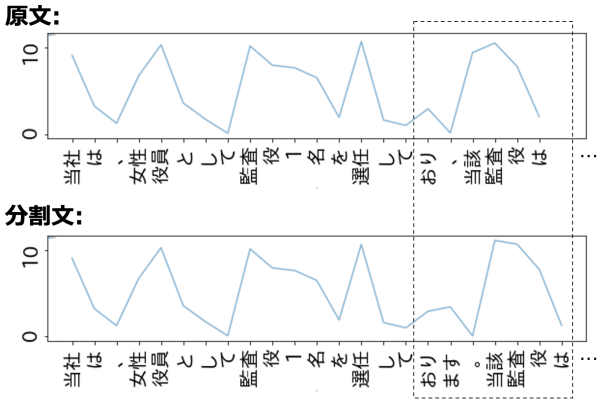

図 3　トークン単位のサプライザルの算出結果の例 (点線での強調は、文分割によってトークンに変化が生じた箇所の付近を示す)

表 4　各指標の評価結果 (いずれも平均値)

評価指標	原文	分割文	差分	増減率
文字列長	150.1	156.5	6.3	+3.0%
トークン数	90.0	98.3	4.3	+4.5%
係り受け木の深さ	50.4	31.9	-18.5	-36.0%
サプライザルの総和	525.7	540.4	14.7	+3.0%

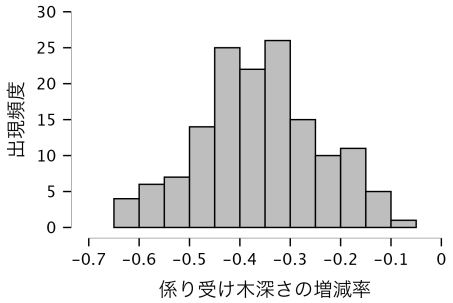

図 4　係り受け木の深さの増減率の分布

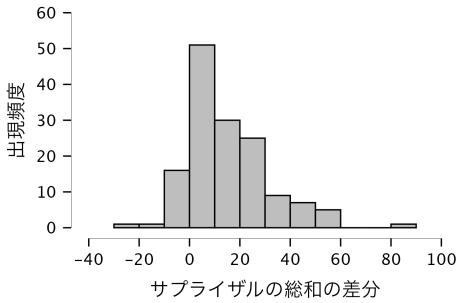

図 5　サプライザルの総和の差分の分布

ン数は少なくなるが、全体としては、文分割と共に行う語句の補完による影響を受け、これらの指標が増加したものと推察する。

係り受け木の深さは、図 4 のとおり、全体的に浅くなっており、増減率の平均値は-36.0%だった。この結果は、長文の文分割によって文構造の把握しやすさが向上することと齟齬は無く、文分割による読みやすさへの影響を定量的に示すものと考える。

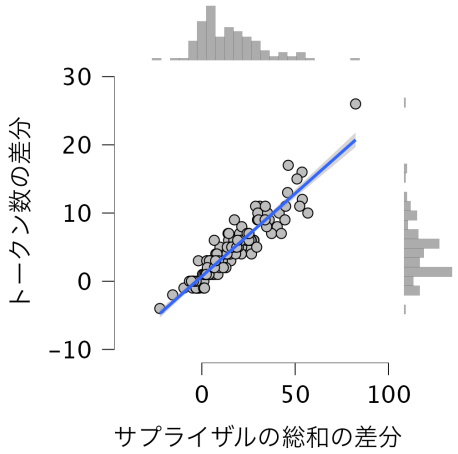

図 6　サプライザルの総和とトークン数の差分の散布図

一方で、サプライザルの総和を見ると、ほとんどの分割文において上昇していることが確認された。図 5 及び図 6 に、サプライザルの総和の差分とトークン数の差分の分布を示す。図 6 より、サプライザルの総和とトークン数には強い正相関が認められており (相関係数: 0.934、p 値: < 0.001)、サプライザルの総和は文分割によるトークン増加の影響を強く受けているものと推察する。すなわち、サプライザルの総和は、文分割によるトークン数の増加によって増加しており、文分割による低減は認められなかった。この結果から、文分割による読みやすさの変化の評価には、各トークンのサプライザルの総和が適しているとは言えなかった。そのため、文全体の総和ではなく、文節やトークンといった、より細かい粒度でのサプライザルを通じた分析が、今後の課題として考えられる。また、今回サプライザルの算出に用いた言語モデルが、開示資料中の文の言語モデルと乖離している可能性もあり、この検証も課題の一つとして挙げられる。

5　おわりに

本研究では、上場会社開示資料中の一種である CG 報告書を対象として、人手による長文の文分割による、読みやすさに関する指標への影響について分析した。実験の結果、長文の文分割によって、係り受け木の深さが平均 36%程度低減することを示した。一方、サプライザルの総和による指標では、文分割による読みやすさへの影響は評価できなかった。本研究は、作成した文分割コーパスは 146 文対と小規模なものであり、試験的な研究であったことから、データセットの拡充等を通じて、文分割に係るより広範かつ詳細な分析を検討したい。

謝辞

本研究において、東京大学先端科学技術研究センターの田中久美子教授に有益なご助言を戴いた。ここに記して謝意を表する。

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**, pp. 3719–3726, Marseille, France, May 2020. European Language Resources Association.

[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. 525–528, 3 2021.

[3] 土井惟成. プリエディット手法としての産業日本語に関する一考察. In **Japio YEAR BOOK 2020**, pp. 324–331, 2020.

[4] 土井惟成. 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. In **Japio YEAR BOOK 2021**, pp. 326–331, 2021.

[5] 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 (第 2 版), 第 2 版, 3 2019.

[6] 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 9 回年次大会 (NLP2013), pp. 264–267, 3 2013.

[7] 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**, pp. 606–616, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 806–817, 2017.

[9] Gunning Robert. The technique of clear writing. **New York**, 1952.

[10] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[11] Sarah Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**, pp. 523–530, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[12] Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. Automatic assessment of Japanese text readability based on a textbook corpus. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[13] Hideko SHIBASAKI and Katsuo TAMAOKA. Constructing a formula to predict school grades 1-9 based on japanese language school textbooks. **Japan Journal of Educational Technology**, Vol. 33, No. 4, pp. 449–458, 2010.

[14] Alan Kennedy, Joël Pynte, Wayne S Murray, and Shirley-Anne Paul. Frequency and predictability effects in the dundee corpus: An eye movement analysis. **Quarterly Journal of Experimental Psychology**, Vol. 66, No. 3, pp. 601–618, 2013.

[15] Masayuki Asahara. Between reading time and information structure. **Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation**, pp. 15–24, 2017.

[16] John Hale. A probabilistic Earley parser as a psycholinguistic model. In **Second Meeting of the North American Chapter of the Association for Computational Linguistics**, 2001.

[17] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, p. 1126–1177, 2008.

[18] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 日本語の読みやすさに対する情報量に基づいた統一的な解釈. 言語処理学会第 27 回年次大会 (NLP2021), pp. 723–728, 3 2021.

[19] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[20] 渡邉亮彦, 村上聡一朗, 宮澤彬, 五島圭一, 柳瀬利彦, 高村大也, 宮尾祐介. TRF: テキストの読みやすさ解析ツール. 言語処理学会第 23 回年次大会発表論文集, pp. 477–480, 2017.

[21] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In **Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing**, pp. 186–195, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[22] Taku. Kudo. Mecab : Yet another part-of-speech and morphological analyzer. **http://mecab.sourceforge.net/**, 2005.