上場会社開示資料の日英対訳コーパスの自動生成に関する検討

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

1 はじめに

東京証券取引所 (以下，東証) における海外投資家のプレゼンスは年々高まっている．東証が公表している株式分布状況調査 [1] によると，2019 年度における日本の上場会社の株式に対する外国法人の保有比率 (時価総額に対する金額ベース) は約 30%に及んでいる．しかしながら，2018 年に TDnet[1] で開示された，適時開示資料をはじめとする上場会社開示資料 (以下，開示資料) において，英語による開示資料の占める割合は約 11 %に留まっている．

このような状況を踏まえ，海外投資家による開示資料の利便性を改善するための手段として，機械翻訳の活用が考えられる [2]．ニューラル機械翻訳モデル [3] (以下，NMT モデル) による機械翻訳は，従来の機械翻訳モデルより翻訳精度が高いことで知られている．一方で，専門用語の多い特定分野の文を対象にすると，汎用的な NMT モデルでは翻訳精度が低くなる傾向にある [4]．また，NMT モデルの学習は対訳コーパスを用いて行われるため，対訳コーパスの規模や品質が翻訳精度に影響を与えることが知られている [5]．

本研究では，2019 年に TDnet で開示された，PDF 形式の開示資料を対象に，約 23 万文対の日英対訳コーパス (東証適時開示対訳コーパス: ParaTDC) を機械的に生成した．その後，ParaTDC の有用性の評価のため，汎用的な NMT モデルを対象に本コーパスを用いて追加学習を行うことでカスタム翻訳モデル[2] を作成し，その翻訳精度を評価した．実験の結果，日英の文の類似度で閾値を設けた対訳コーパスを用いた追加学習によって，追加学習前と比較して BLEU スコア [6] の上昇が確認できた．

1） https://www.release.tdnet.info/inbs/I_main_00.html
2） 分野適応モデル，カスタムモデルとも呼ばれる．

2 関連研究

本研究に関連する研究として，(1) 開示資料の一つである CG 報告書を対象としたカスタム翻訳モデルの検討に関する研究 [7]，(2) 開示資料を対象とした人手による日英対訳コーパス (TDDC) の作成に関する研究 [8]，(3) Web サイトのクロールで収集した HTML ファイルを対象とした日英対訳コーパス (JParaCrawl) の自動生成に関する研究 [9] がある．以下では，これらについてそれぞれ紹介する．

(1) の研究では，160 社の日英の CG 報告書を対象に，約 4 万文対の対訳コーパスを人手で作成し，カスタム翻訳モデルを評価した．本研究と (1) の主な差異は，(1) の対象とする開示資料が限定的である点と，作成手法が人手である点の 2 点が挙げられる．

(2) の研究では，2016 年 1 月から 2018 年 6 月までの 2.5 年間に TDnet で開示された PDF 形式の開示資料を対象に，人手によって約 140 万文対の日英対訳コーパス (TDDC) を作成した．本研究と (2) の主な差異は，(2) の対象とする文書の開示期間と，作成手法が人手である点の 2 点が挙げられる．

(3) の研究では，Web クロールで収集した HTML ファイルを対象に，機械的な手法で作成された，約 1,000 万文対の日英対訳コーパス (JParaCrawl) を作成した．ParaTDC と JParaCrawl は，大量の日英の文書から自動的に作成した日英対訳コーパスという点で共通している．一方で，本研究と (3) の主な差異は，(3) の対象とする文書が HTML ファイルである点と，文単位のアライメント手法に内製の NMT モデルを用いている点の 2 点が挙げられる．

3 対訳コーパス生成の流れ

2019 年における TDnet の開示資料の件数は約 10.3 万件であり，その内，日本語は 9.2 万件，英語は約 1.1 万件である．本研究では，これらの開示資料を源泉として以下の各節に述べる処理を実行した．

3.1 暗号化・画像化された PDF の除外

PDF ファイルからのテキスト抽出を防ぐ方法として，コピー禁止の設定，パスワードや電子署名による PDF 文書の暗号化，テキストの画像化が挙げられる．手書き文字認識等の技術を用いることで，これらの処理が施されている PDF ファイルからのテキスト抽出が可能ではあるものの，その精度は必ずしも高いとは限らない．そのため，本研究では，これらの処理が施されている PDF ファイルは，コーパスの源泉データの対象から除外した．

3.2 日英の文書のアライメント

対訳コーパスの源泉となる開示資料に対して，文書単位での日英のアライメントを実施した．TDnet の開示資料には，PDF 形式の開示資料本体とは別に，開示日時やタイトル等のメタデータが存在する．このメタデータから，開示資料の言語 (日英) は判別できるが，日英の対応関係が含まれていないため，別途の手段による日英の文書のアライメントが必要となる．また，英語の開示資料は，必ずしも対応する日本語の開示資料と同日に開示されるとは限らず，日本語より遅延して開示されることがある．

本研究では，まず，英語の各開示資料に対して，2019 年 1 月 1 日から各開示日までに，その銘柄コードの上場会社が開示した全ての日本語の開示資料を出力した．その後，日英のタイトルの分散表現を算出し，英語の開示資料のベクトルとのコサイン類似度が最大となる日本語の開示資料を探索することで，文書のアライメントを実現した．この時，テキストの分散表現の算出に TensorFlow の Embedding メソッド[3]を使用し，分散表現のモデルには，TensorFlow の universal-sentence-encoder-multilingual[4] を使用した．

3.3 PDF テキスト抽出

PDF ファイルには，セクション，見出し，表，段落などの論理的な構造を追加するためのタグを付与することが可能であり，このようなタグが付与された PDF ファイルをタグ付き PDF と呼ぶ．しかしながら，開示資料におけるタグ付き PDF の割合は一定程度に留まっており (2018 年の開示資料では 28%程度)，また，タグ付き PDF であっても，必ずしも全てのタグが適切に付与されているとは限らない．

そこで，本研究では，PDF 形式の開示資料からのテキスト抽出には，文書中の各文字の座標情報等を踏まえて，一連のテキスト (以下，テキストボックス) を抽出することが可能なソフトウェアを利用し，テキストボックス単位でテキストを抽出した．具体的には，PDF ファイル中の文字について，文字間隔が閾値以下同士の文字は，同一の行に含まれる一連のテキスト (以下，テキストライン) に属すと見なし，行間隔が閾値以下同士のテキストライン同士は同一のテキストボックスに属すと見なしている．このような処理を有するソフトウェアとして，PDFMiner[10] が知られている．なお，本処理では，文書中の 1 パラグラフが 1 つのテキストボックスとして抽出されるため，1 つのテキストボックスに 2 文以上が含まれることがある．

3.4 正規化処理

開示資料から抽出したテキストを対象に，以下をはじめとする正規化処理を実行した．

特定文字の置換　長音記号や全角チルダをはじめとする一部の記号文字を，類似する形状の記号文字に置換した．また，CJK 部首補助や CJK 互換漢字は後述する NFKC 正規化によって正規化されないことから，本処理においてそれぞれ対応する CJK 統合漢字へ置換した．

Unicode 正規化　丸数字 (U+2460 - U+2473)，2 点リーダ (U+2025)，3 点リーダ (U+2026) を除く文字を対象に，NFKC[11] による Unicode 正規化を実行した．これにより，全角英数字を半角に変換したほか，康熙部首を CJK 統合漢字に変換した．康熙部首とは，部首を表す漢字の一種であり，一部の PDF 変換ソフトウェアにおいて，CJK 統合漢字で入力した文字が PDF ファイルでは康熙部首に変換されることが知られている [12].

制御文字等の削除　制御文字に相当する文字等を削除する目的で，Unicode character property の General Category[13] が Cc，Cf，Cn，Co の文字を削除した．

余分なスペースの削除　以下の処理を行うことで，余分なスペースを削除した．

- 文頭及び文末のスペースを削除
- 連続するスペースを 1 つのスペースに置換
- 「平仮名，片仮名，CJK 統合漢字，全角記号」の間のスペースを削除

3) https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding
4) https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

- 「平仮名，片仮名」と「半角濁点 (U+FF9E)，半角半濁点 (U+FF9F) 」の間のスペースを削除し，NFKC 正規化を実行

3.5 日英の文のアライメント

アライメントされた日英の文書ごとに，日英の文のアライメントを行った．本処理では，まず，和文と英文のそれぞれに対して分散表現を算出した．この時使用したモデルは，3.2 節におけるタイトルの分散表現の算出に使用したモデルと同様である．この時，日本語側の文については，日本語の文字を含まない文 (平仮名，片仮名，CJK 統合漢字のいずれも含まない文) は処理の対象から除外した．その後，英文ごとに，英文のベクトルとのコサイン類似度が最大となる和文のベクトルを探索し，それを対訳文として選定した．

この時，文書全体での類似度の組み合わせの最適化等は行っていない．そのため，文書中に複数出現する同一の和文に対して異なる英文がアライメントされることや，英文がアライメントされない和文が生じることがある．

また，本処理の実行に際しては，計算コスト上の理由から，1 文書当たり 10 分を処理時間の上限として設け，それ以降の処理を中断した．

4 ParaTDC の概要

本研究では，3 章の処理を通じて，2019 年の開示資料を源泉とした，約 23 万文対の日英対訳コーパスである ParaTDC を生成した．ParaTDC における対訳数や平均文字列長等の詳細を表 1 に示し，英文における単語数の分布を図 1 に示す．

TDDC が 2.5 年分の開示資料を源泉とした 140 万文対の対訳コーパスであることを踏まえると，ParaTDC は開示資料 1 年分当たりの対訳数が比較的少ないと言える．この理由としては，3.3 節の PDF テキスト抽出において，1 行中に複数の文を含むことを許容していること，また，3.4 節の文アライメントにおいて，10 分以上の処理を途中で中断していることが考えられる．

また，ParaTDC は，日英の対訳に加え，源泉の開示資料の ID，開示日時，発行体の銘柄コード，3.4 節で算出したコサイン類似度 (以下，スコア) で構成されている．スコアは必ずしも翻訳の正確性を表しているとは限らないものの，対訳コーパスからのノイズ除去において有用性が期待される．

表 1 ParaTDC の統計情報．括弧内の数値は全対訳数に占める割合を示す．

項目	値
全対訳数	236,653 文対
抽出元の書類数 (日英のペアの数)	8,011 文書
和文 100 文字以上の対訳数	23,540 文対 (9.90%)
英単語 50 単語以上の対訳数	17,007 文対 (7.20%)
1 文当たり平均文字列長 (和文)	48.8 文字
1 文当たり平均英単語数 (英文)	20.0 単語

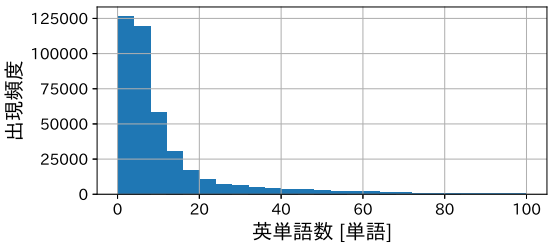

図 1 ParaTDC の英文における単語数の分布

5 実験

ParaTDC の有用性を検証するため，ParaTDC と TDDC (2016 年及び 2017 年の開示資料) のそれぞれを用いてカスタム翻訳モデルを作成し，同一のテストデータを用いて翻訳精度の向上を検証する．5.1 節では本実験の実験設定について述べ，5.2 節では実験結果と考察について述べる．

5.1 実験設定

本実験では，汎用的な NMT モデルに追加学習を行うことでカスタム翻訳モデルを作成するサービスとして，Microsoft が提供する Custom Translator[5] を利用した．また，翻訳精度の評価指標として BLEU スコアを採択した．

ParaTDC を用いた追加学習では，Dev には訓練データからランダムに選出した 5%程度の対訳 (上限 2,500 文対)，Test には TDDC で提供している，2018 年 1 月から同年 6 月の開示資料から作成した 3,277 文対を用いた．TDDC を用いた追加学習では，Train には 2016 年と 2017 年の開示資料から作成した対訳コーパス，Dev には 2018 年 1 月から同年 6 月の開示資料から作成した 3,967 文対，Test には ParaTDC と同一の 3,277 文対を用いた．この時，Dev と Test は，作成元の開示資料がそれぞれ異なることを確認している．なお，Custom Translator の仕様として，英単

5) https://portal.customtranslator.azure.ai/

表 2 実験結果．閾値を設けている場合，スコアが閾値未満の対訳を元の対訳コーパスから除外していることを意味する．

評価対象	閾値	使用した対訳数		BLEU スコア	
		Train	Dev	追加学習	上昇量
ParaTDC	なし	174,886	2,500	29.29	1.82
	20%	172,600	2,500	29.50	2.03
	40%	153,228	2,500	29.78	2.31
	60%	94,133	2,500	30.22	2.75
	80%	17,830	938	25.83	-1.64
TDDC	-	601,832	3,967	36.35	8.88

語数が 100 個以上の対訳や，和文が 2,000 文字以上の対訳は自動的に削除される [14]．

ParaTDC 及び TDDC には重複する対訳が含まれており，追加学習に際してはこれらを除外した．また，ParaTDC に重複して出現する和文については，スコアが最も大きい英文のみを残すことで，各和文に対して対応する英文が一意に定まるようにした．

また，本実験では，ParaTDC のスコアによるノイズ除去の有用性の検討のため，複数のパターンでスコアに閾値を設けた対訳コーパスを用いてカスタム翻訳モデルを作成し，翻訳精度を評価した．

5.2 実験結果及び考察

各追加学習で用いたデータセットの内訳や BLEU スコアによる評価結果を表 2 に示す．

ParaTDC による追加学習により，多くの閾値のパターンにおいて BLEU スコアの上昇が見られた．特に，スコアによる閾値を 60%とした時の上昇量が最も大きく，2.7 ポイント程度の上昇を確認した．

一方で，TDDC では BLEU スコアが 8.8 ポイント程度と，ParaTDC と比較すると大幅に上昇した．ParaTDC による翻訳精度の改善が TDDC に及ばなかった原因の一つに，対訳コーパスの規模が考えられる．4 章で述べたとおり，ParaTDC の作成に当たっては，1 文中に複数の文が含まれる可能性があり，また，文書中の全ての文に対してアライメントを行っているとは限らないため，今後はこれらの解決が課題となり得る．

スコアによる閾値に着目すると，閾値を 80%にした時の BLEU スコアはベースラインよりも悪化した．これは，閾値を上げることによって追加学習に用いる学習データが減ったことに加え，アライメントのスコアの高い対訳が，必ずしも正しい対訳であるとは限らないことに起因していると推察する．表 3 に，ParaTDC におけるスコアの高い対訳の例を示す．ParaTDC では，表 3 の例 1 のとおり，短文のアライメントは比較的正確であり，そのスコアは高い傾向にある．一方で，個人名 (例 2) や施設名 (例 3) をはじめとする固有名詞，及び，金額 (例 4) や日付 (例 5) をはじめとする数値表現において，アライメントが誤っているにもかかわらず高いスコアを示している対訳が散見された．この原因として，固有名詞や数値表現の分散表現では，それらが異なる値であるにもかかわらず，ベクトルとしては似通ってしまうことが挙げられる．これを回避する方法として，固有表現抽出の利活用が考えられる．

表 3 アライメントのスコアが高い対訳の例

例	和文	英文	スコア
1	2019 年 GRESB 評価の詳細については、GRESB のウェブサイト (https://gresb.com) をご参照ください。	For details of the 2019 GRESB assess–ment, please refer to the GRESB website (https://gresb.com).	88.9%
2	【個人の氏名】	2. About GRESB	89.1%
3	IMP 守谷 2	February 28, 2019	81.2%
4	936,227 千円 (注 2)	2,190 million yen	89.6%
5	2019 年度累計	February 7, 2019	80.0%

6 おわりに

本研究では，PDF 形式の上場会社開示資料から機械的に対訳コーパスを生成し，それを用いた追加学習によるカスタム翻訳モデルの翻訳精度を評価することで，ParaTDC の有用性を評価した．実験から，ParaTDC による追加学習によって翻訳品質の改善を確認した．一方で，対訳コーパスの規模及び品質には改善の余地があると考えられる．

東証上場会社は毎年膨大な量の開示資料を開示していることから，引き続き，これらの分析を進めることで，自然言語処理による開示資料の利活用の可能性について検証を進めたい．

謝辞

ParaTDC の作成に当たっては，東証が 2020 年 4 月から同年 8 月に実施した，東証適時開示コーパスに関する限定公開実証実験[6] の参加者各位より有益なご助言を戴いた．ここに記して謝意を表する．

6) https://www.jpx.co.jp/corporate/news/news-releases/0060/20200416-01.html

参考文献

[1] 株式会社東京証券取引所. 2019 年度株式分布状況調査, 2020. https://www.jpx.co.jp/markets/statistics-equities/examination/01.html.

[2] 山藤敦史. 開示資料と翻訳. AAMT 2019,Tokyo 〜機械翻訳最前線〜, 2019.

[3] 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.

[4] 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.

[5] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 211–221, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pp. 311–318, 2002.

[7] 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.

[8] 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.

[9] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[10] Yusuke Shinyama. PDFMiner, 2016. https://euske.github.io/pdfminer/.

[11] Ken Whistler. UAX #15: UNICODE NORMALIZATION FORMS, 2020. https://www.unicode.org/reports/tr44/.

[12] Create PDF, why KANJI 9AD8(高) will be changed to 2FBC(高) when Meiryo UI ? - Adobe Support Community - 10625575, （2021-01 閲覧）. https://community.adobe.com/t5/acrobat/create-pdf-why-kanji-9ad8-%E9%AB%98-will-be-changed-to-2fbc-when-meiryo-ui/td-p/10625575?page=1.

[13] Ken Whistler. UAX #44: UNICODE CHARACTER DATABASE, 2020. https://www.unicode.org/reports/tr15/.

[14] データのフィルター処理 - カスタム トランスレーター - Azure Cognitive Services | Microsoft Docs,（2021-01 閲覧）. https://docs.microsoft.com/ja-jp/azure/cognitive-services/translator/custom-translator/data-filtering.