# 汎用言語モデル「BERT」のビジネス実用化技術に迫る

**ことばを処理するAI（人工知能）技術の1つとしてBERTという技術が脚光を浴びています．NTTデータではBERTをビジネスで実用化するため，さまざまな業界の特有な単語や言い回しをとらえることのできるBERTの応用開発を行っています．これにより，個々のお客さまニーズに合わせた最適なAIモデルを構築することができます．本稿では，その一例である「金融版BERT」と「ドメイン特化BERTフレームワーク」について紹介します．**

和知（わち）　徳磨（とくま）

NTTデータ

## はじめに

近年，深層学習をはじめとしたAI（人工知能）関連技術がめざましい進歩を遂げており，これまで実用足り得なかった分野に対して適用が進んでいます．画像処理では，犬と猫の区別も難しかった状態であったのに対し，深層学習を活用することで犬種の区別やピクセルレベルでの位置まで特定することができるようになりました．言語処理では，機械翻訳が10年前と比べて劇的に改善されています．個々のタスクでは「この文の著者は否定的な見解を述べている」や「この単語は人名を表している」というようなことが機械的に判別できるようなっています．

自然言語*1処理を支える技術としてBERT（Bidirectional Encoder Representations from Transformers）という技術が大きなブレイクスルーを引き起こし，注目を浴びています．しかし，BERTをビジネス適用する際の課題として，金融や医療などの業界特有の専門用語や専門知識が多く含まれた文書に対して，期待する精度が実現できないことがあります．本稿ではそれを解決する技術である金融版BERTおよびドメイン特化BERT技術について紹介します．

## 自然言語処理技術

深層学習の技術は，分類や検出，数値予測，生成などの幅広いタスクで高い精度を実現し，一部のタスクでは人間を上回る精度を達成しています．2015年には，画像分類のベンチマークタスクで人間を上回る精度を達成して話題になりました．

自然言語処理の分野も発展してきており，従来のパターン認識や出現頻度をベースとした手法から，深層学習をベースとした手法へシフトしています．これにより以下のようなこ

＊1　自然言語：「日本語」，「英語」などの人間が日常的に用いている言語．

とが機械的に実現できるようになっています．

・「この文の著者は否定的な見解を述べている」（ネガポジ判定）
・「文章を入力として，記載内容から潜在リスク等を数値化」（スコアリング）
・「文中から人名・地名のような特定の種類の単語を抜き出す」（固有表現抽出）
・「文章を読み込み，文章に対する質問に回答する」（質問応答）

自然言語処理の分野では，単語や文の単位で入力を処理することが一般的です．最近では，一般的な文章に対して単語や文を処理する汎用的なモデルを用意して，このモデルを各タスクに合わせてチューニングする方法がよく用いられています．この汎用モデルは「言語モデル」と呼ばれています．後述するBERTも言語モデルの１種です．

## BERTとは

BERTはGoogleが開発した汎用自然言語処理モデルです．2018年に発表された際に，さまざまな自然言語処理のベンチマークタスクの従来記録を塗り替えたことで話題になりました．例えば，「Wikipediaから抜き出した140単語程度の文を読み内容に関する質問に答える」というようなタスクがベンチマークとして採用されており，人間より高い精度を実現しました．

BERTの強みはさまざまなドメインやタスクの課題を単一のモデルで解くことができる点です．BERT以前のタスク特化型モデルでは，解きたい課題に対して言語自体の特徴を含めゼロから学習を始めるため，課題ごとに多数の教師*2データが必要でした．BERTは大規模な文章群による教師なし事前学習を行うことで，教師データを必要とせずに汎用的なモデルをつくり上げています．このように，大量の自然文をそのまま使うだけで汎用言語モデルを構築できる仕組みをつくり上げたことが，BERTの功績といえます．しかし，このような汎用事前学習モデルをつくることができるのは，技術力や計算資源が豊富な一部の団体に限られているのが現状です．

そしてBERT利用時は事前学習済BERTモデルを少量の教師データでチューニングすることで，対象の課題に対して高精度を実現できるといわれています．例えば，「著者の見解を肯定と否定に分類するタスク」の場合，言語モデルの後段に小規模な重み付けモデルを追加し，肯定度合いと否定度合いをそれぞれ数字で出力するようにして，その大きさを比べて最終結果を出力する，といったチューニングを行うことが一般的です．

## BERTの日本語化とNTT版BERT

Googleが発表したBERTは英語版でしたが，日本では京都大学や情報通信研究機構（NICT）などが日本語版の事前学習済モデルを公開しています．BERTの汎用事前学習モ

*2　教師（データ）：解きたいタスクに合わせて，データにAIモデルが解釈できるようなラベル付けを行うこと（またはラベル付けされたデータ）．例えば，ある商品のレビュー文を「好意的」・「非好意的」で分類することや，文章中に出てくる「人名」・「地名」部分に目印をつけることを行います．また，データに教師情報を付与することをアノテーションともいいます．

デル構築の肝は，事前学習で用いる文書群の質（多様性）と量をいかに担保するか，にあります．当初，日本語版の事前学習モデル構築では，一定の質と量が担保され，かつ入手が容易である日本語Wikipedia全文（約3 GB）を用いる方法が主流でした．しかしこの方法では，Wikipediaで出現頻度が少ない文体（話し言葉など）に対して，性能が発揮しづらいという問題点も分かってきています．

NTT研究所では，独自に収集した大規模文章群（約13 GB）から日本語の事前学習モデルを構築しており，多くのタスクで公開されている事前学習済モデルより高い性能を発揮しています．後述するBERTも特に断りのない限り，NTT研究所のBERTを基にしています．

## ドメイン特化のための追加学習

BERTやその日本語版は従来技術に比して高い性能を実現しています．例えば，金融分野では「FAQの回答自動引き当て」や「財務情報からのリスク抽出」，医療分野では「電子カルテの記載内容チェック」や「医薬品添付文書の情報活用」のようなユースケースでの活用が期待されています．しかし，前述の大規模な一般文章群で事前学習を行った汎用モデルをチューニングするやり方では，実際のビジネス適用において期待するほどの精度が実現できないこともあり，課題となっています．

この傾向はビジネス課題の対象となるデータが特定のドメインに偏っている場合（ドメインデータ）で特によくみられます．例えば，「金融や医療などの専門語を多く含んだデータ」や「道路交通法や慣習などの特有の知識が必要な運転関連データ」などがドメインデータの例といえます．

ドメインデータごとに大規模文書群を用意することは現実的ではなく，特定のドメインタスクで汎用BERTより精度向上を図りたい場合，何らかの工夫をする必要があります．

最近の研究では，小〜中規模の文章群を事前学習済BERTモデルに追加学習を行うことにより，ドメインに適応した事前学習モデルを構築する手法が提唱されています（**表**）．つまりタスクに特化した言語モデルをつくるということです．NTTデータが開発中である，金融版BERTやドメイン特化BERTでも同様の追加事前学習アプローチを採用しています．

表　データ規模の違い

※データ規模の考え方は，対象としている問題や文脈によってさまざまですが，本稿では以下に基づくものとします．

| 大規模 | 中規模 | 小規模 |
|---|---|---|
| 人手では全体の傾向をとらえることも難しい規模の文章群 | 人手では流し読みでも全件確認は難しい規模の文章群 | 人手で全件確認や教師データ付与が可能な規模 |
| 数GB以上 | 数MB〜数百MB | 数KB〜数百KB |

## 金融版BERTの仕組み

NTTデータではインターネット上から収集した金融関連文章を用いて追加事前学習を行い，金融ドメインに特化した事前学習済モデルである金融版BERTを構築しました．金融版BERTの検証として一種外務員資格試験[(1)]の解答に取り組み，検証したモデル中で唯一合格点相当のスコア（440点満点中308点）を実現しました．

外務員資格試験とは日本証券業協会が主催している試験で，証券取引・デリバティブ取引の勧誘を行う外務員が受験する試験です．その中でも一種試験は上位資格であり，2019年に4633名の受験者が受験していて，合格率は67.6%です．試験の内容は，主に問題文の正誤を判定する○×問題と，正しい文を選ぶ5択の選択問題で構成されます．問題文中には金融商品の専門用語や法令に関する知識が登場し，一般的な知識のみでは正解が難しい問題も多くあります（**図1**）．

ここでは金融版BERTの仕組みについて説明します（**図2**）．

① 金融関連文書をインターネット上から収集（NTTデータが実施した金融分野へのAI適用案件において蓄積した知見を基に，効率良くBERTモデルの性能向上を行えるようなWebページを選択）

正解：×

金融版：協会員は，日本証券業協会の審査により「二級不都合行為者」とされた者については，その決定を受けた日から3年間はいかなる名称を用いているかを問わず採用してはならないこととされている．

NTT版：協会員は，日本証券業協会の審査により「二級不都合行為者」とされた者については，その決定を受けた日から3年間はいかなる名称を用いているかを問わず採用してはならないこととされている．

正解：×

金融版：日経225先物には，制限値幅は定められていない．

NTT版：日経225先物には，制限値幅は定められていない．

モデルが着目している個所（濃いほど重要と判断）

赤字 金融版BERTのみが着目している個所

図1　金融版BERTで正答できるようになった問題例

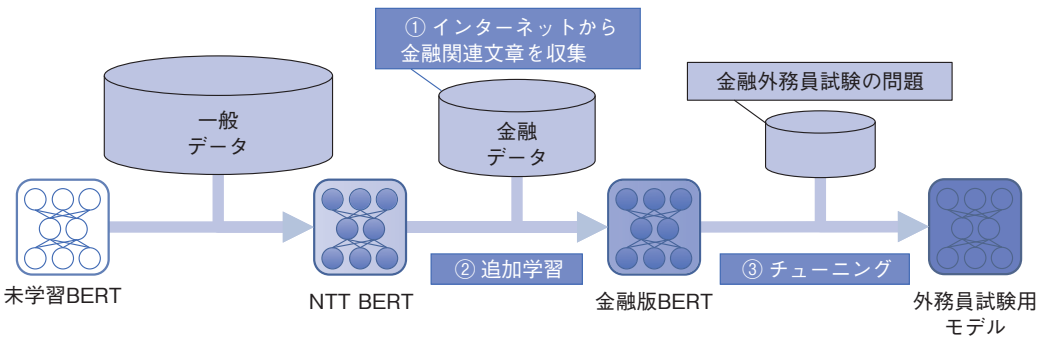


図2　金融版BERTの処理概要

② NTTが開発した日本語BERTを基に，上記の文章群を追加事前学習

③ ビジネス課題を反映したタスクデータを用いたチューニング

①～②について，NTTデータが事前学習済モデルを保有しているため，実際のPoC（Proof of Concept）やシステム開発への適用には，ビジネス課題に応じた教師付タスクデータの作成とそのデータに応じたチューニングのみが必要となります．

## ドメイン特化BERTフレームワークの仕組み（開発中）

金融版BERTでは，汎用BERTモデルをドメイン特化させることで高精度を実現することができました．一方で，金融関連文書を収集する際には有識者が対象とするWebページを選定してデータを収集していました．このため，モデル構築の際のコストが高く，また有識者の参画が必須である，という課題があります．これらの課題を解決するため，NTTデータでは追加学習用のドメインデータ収集の自動化をめざしたドメイン特化BERTフレームワークを開発しています．

ここではドメイン特化BERTフレームワークの仕組みについて説明します（**図3**）．

① ビジネス課題を反映したタスクデータを準備（この時点では教師なしでも良い）

② タスクデータに応じた追加事前学習用の文章をインターネット上から自動収集（タスクデータ中から汎用BERTが苦手としている表現を抽出し，その情報からクエリを生成してインターネット検索を実施）

③ 収集したデータから追加事前学習で精度向上が期待できる文章をアルゴリズムで選別（インターネットから収集した文章群から，タスク精度向上が期待できる文を抽出してドメインデータのデータセットを作成）

④ NTTが開発した日本語BERTに，上記の文章群を追加事前学習

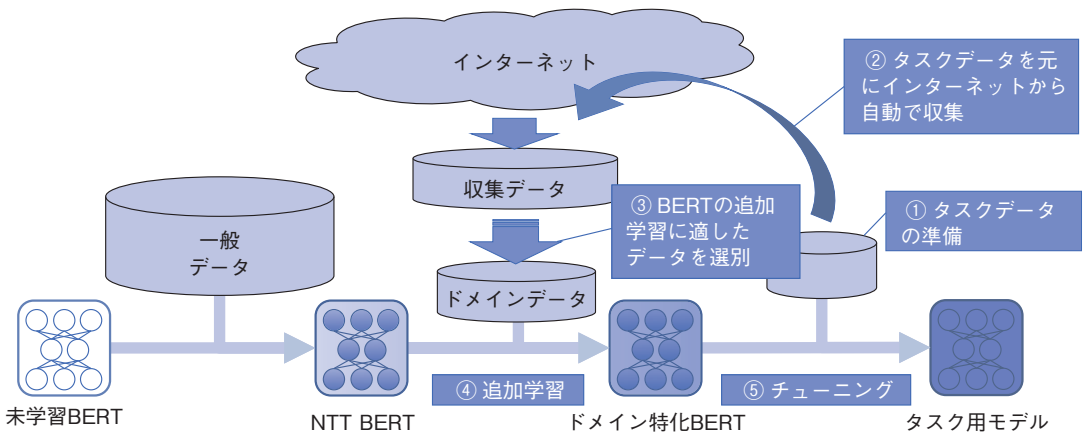


図3　ドメイン特化BERTの処理概要

⑤　タスクデータを用いたチューニング（この時点で教師データが必要）実施

実際のPoCやシステム開発への適用には，金融版BERTと同様に，ビジネス課題に応じた教師付タスクデータの作成とそのデータに応じたチューニングが必要となります．加えてタスクに合わせた追加事前学習を自動化された基盤上で行うことで，さらなる高精度をめざします．

ドメイン特化BERTフレームワークの強みは自動化されたデータ収集とデータ選別によりお客さまのデータに合わせた最適モデル構築が行えることと，ドメイン特化BERTを人手で構築する場合と比べてPoCや開発期間を短縮できることです．BERT以前の技術では，専門用語を多く含むドメインに特化したタスクを扱う際には，タスクごとに人手で辞書をつくるなどの個別対応が必要でした．BERTを用いることで辞書構築の代わりに大量の文書群を用いることで汎用的な言語モデルを構築し，それを活用することでさまざまなタスクに対する改善を実現しました．さらにドメイン特化BERTフレームワークを用いることで，特定のドメインタスクに対するさらなる精度向上を期待できます．

## おわりに

本稿では，BERTおよびその日本語化，それを応用した金融版BERT・ドメイン特化BERTフレームワークについて紹介しました．現在開発中であるドメイン特化BERTについては，さらなる精度向上と効率化へ向けて改良を重ねています．

2021年度以降，金融・医療・製造などのお客さまに提供し，新規ビジネスの創出や既存ビジネスの効率化にお力添えさせていただきます．例えば，「FAQの回答自動引き当て」「電子カルテの記載内容チェック」「日報からのプロジェクトリスクチェック」等の活用ユースケースが例に挙げられます．

また，前述の業界・事例に限らず，BERTの先進的な技術をいち早くビジネス適用できるようお客さまを支援していくとともに，ドメイン特化BERTフレームワークを活用したPoCパートナーを募集中です．

**■参考文献**

⑴ https://www.jsda.or.jp/gaimuin/shiken.html

**和知　徳磨**

NTTデータでは，金融版BERT・ドメイン特化BERTをはじめとしたAI関連技術の開発を行い，ビジネスにおける効率化や新しい価値の提供を行っています．AIや言語処理に関する困りごとがありましたら，ご相談いただけると幸いです．

**◆問い合わせ先**

NTTデータ
技術革新統括本部　技術開発本部
TEL　050-5546-9741
FAX　03-3532-0488
E-mail　ai-coe-jp＠kits.nttdata.co.jp