
「仕事のAI」における大規模言語モデルの個別チューニングを可能にするデータとモデルの共同設計

Joint Design of Data and Model for Large Scale Language Model Tailoring in “AI for Your Biz”

鈴木 剛*
Takeshi SUZUKI

金崎 克己*
Katsumi KANASAKI

要 旨

BERTに代表される近年の言語モデルは、事前学習—ファインチューニングフレームワーク（Pretraining-Finetuning Framework）に基づく表現学習とアプリケーション学習の分離、及び大量のコーパスを表現学習に活用する手段の確立により、多くのタスクで高い精度を実現している。しかし実際のアプリケーション構築では、ドメイン固有の語彙や言い回しなどを言語モデル層及びアプリケーション層の両方で獲得する必要があり、このためファインチューニングでは大量のアノテーション付きデータを必要とする。

本稿で提案するアプローチでは、ファインチューニングに少量のデータしか利用できない実際の応用事例に対応するために、識別カテゴリ構造化やアノテーション対象データ選定などの学習データの設計と、マルチタスク学習による頑健な表現の獲得やエポックごとのデータバランス最適化などのアプリケーションの学習戦略を組み合わせることで、必要データ量を縮減しつつアプリケーションの性能を向上させた。

ABSTRACT

Recent language models, e.g., BERT (Devlin, 2018) are characterized by the pretraining-finetuning framework, which separates the representation learning from application learning to utilize the availability of large-scale non-annotated document corpora and improve performance on many different NLP tasks. However, to develop an application with such language models, a huge amount of annotated data is required in finetuning to encode decision logics as well as application domain-specific vocabulary and expressions in both the language and application layers of the model.

In this report, we present a joint design of a training data scheme and an application training strategy, which improves the performance of downstream tasks assuming a limited amount of annotated data in real scenarios. The approach minimizes the amount of required data by combining 1) the training data scheme design and development, i.e., structuring classification categories, and 2) the application training strategy, i.e., multitask learning and pseudo-ensemble training by balancing the training data set in each epoch.

* リコーデジタル戦略部 デジタル技術開発センター DDX開発室
DDX Development Department, Digital Technology Development Center, Digital Strategy Division

1. はじめに

2023年現在、自動車の運転支援や工作機械の制御など様々な分野でAIが導入されている。自然言語処理の分野でも、人の書いた文書の内容をAIが的確に判断したり、人が書いたような文書をAIが生成するなど、新たな技術開発が高速に進んでいる。これらの技術はこれまで人しかできなかった業務上の判断をAIが支援・代行できる可能性があることから、産業に大きな影響を与えるものである。適用対象が業務であることから、AIを前提とした業務の最適化、所謂デジタルトランスフォーメーション（DX）の視点が重要となる。

リコーはこれまでデジタル複合機（Multi-Functional Peripheral: MFP）を中心とした文書画像処理技術や、ドキュメントマネジメント機能を開発してきた。最新のAIの利点をお客様に提供するために、これらの技術、機能による自動処理に加え、業務上の判断を支援するAI機能をもとにしたワークフローを提案する「仕事のAI」サービスを立ち上げた。

近年のディープラーニングに基づく自然言語処理では、特徴量設計（Feature Engineering）は自動化されているものの、モデルの学習に大量のデータを必要とする。そのためこれらの技術を事業で活用するためにはモデル学習を効率的に行うためのデータ開発が重要となる。さらに現実の業務では、例えば抽出すべき事例のデータ数が少ないインバランスデータでの安定した学習など、業務での価値を最大化するためのチューニングも課題となっている。

本稿ではディープラーニングに基づく自然言語処理を実際の業務で活用するための顧客企業のDXに貢献する機能設計、モデル設計、及びデータ開発について紹介する。本節以降、第2節で近年のAI技術及びDXについてまとめ、第3節で「仕事のAI」のサービス、機能、及び技術を紹介する。第4節はリコーコンタクトセンターにおける評価結果、第5節はまとめである。

2. AIとDX

2-1 AI

2010年台のディープラーニングの産業応用に始まった第三次AIブームは、これまでの2回のブームと異なり機械学習を用いたデータの機能化の側面が強い。ディープラーニングは物体認識などのサブシンボリックタスクに対しては特徴量設計を自動化することで性能向上及び開発の効率化に非常に有効だが、自然言語処理などのシンボリックタスクに対して直接適用することができなかった。Salakhutdinov, Hinton¹⁾及びMikolov²⁾が提案した自然言語処理におけるエンベディング技術により、テキストなどのシンボルを連続値ベクトルとして表現することで、ディープラーニング系の技術を自然言語処理に対して適用することが可能になった。

第三次AIブーム開始から現在までの間に、AI開発ツール（IDE）、標準的なデータ（LOD, WikiDataなど）、アルゴリズム（ArXiv）、及びソースコード（GitHub）はオープン且つ共有される傾向が顕著になってきた。このため既に標準的なAI機能は開発リソースがあれば誰でも作れる環境ができており、AI関連のサービスを展開する企業にとっては単なる技術だけでは自らの優位性を確立することが難しくなっている。第三次AIブームの特性を考慮し、この優位性をデータのユニークさに求める企業も多い。

機械学習に基づくAIの性能は特徴の選定に大きく依存し、またこれらの特徴をディープラーニング系のモデルとして学習するためには大量のデータを必要とする。そのためAIサービスの開発では学習データの確保と、特徴の学習に適したデータ構造の設計が課題となっている。

2-2 デジタルトランスフォーメーション(DX)

これまでのデジタル化（Digitization及びDigitalization）がアナログデータのデジタル化や、個別のプロセスのデジタル化を目的としたもので

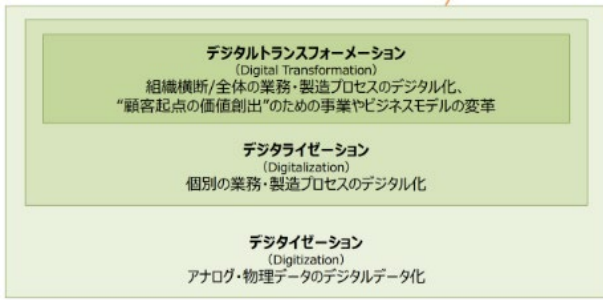


Fig. 1 Digital-Transformation (DX) concept (Japanese Ministry of Economy, Trade and Industry).

あったのに対し、近年話題となっているデジタルトランスフォーメーション（Digital Transformation: DX）は、組織横断／全体の業務・製造プロセスのデジタル化、及び顧客起点の価値創出のための事業やビジネスモデルの変革を目指したものであり、これら3つのコンセプトの間には包含関係がある（Fig. 1, 参考文献3）ため、取り扱うデータ及び個別プロセスのデジタル化を抜きにしてDXを実現することは難しい。またDXには、「破壊的DX」と呼ばれるものから、上記の経済産業省の定義にあるような企業内での活動としてのDXまで幅がある。前者は業界の再編成としての側面が強く企業単位での活動につながりづらいため、本稿では後者の定義に準拠して説明を行う。

一旦データのデジタル化と個別プロセスのデジタル化ができてしまえば、組織横断／全体のプロセスのデジタル化により、ビジネスプロセスで流通する

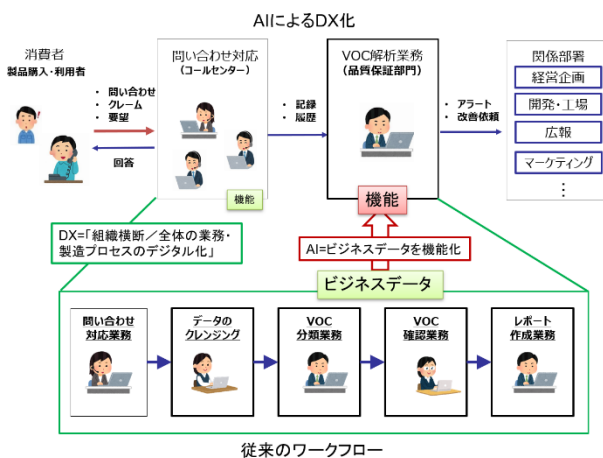


Fig. 2 DX and AI.

データを機能に変換するAIによるアプローチで自動化や効率化など様々な利益に向けたDX化活動を始めることができる。

企業における判断・意思決定などでは客観性や記録としての性質が求められるため、文章による記述や説明が必要不可欠である。そのためAIなどを活用し業務を高度に支援するシステムは、自然言語で記述された業務文書を、入力の一つとして想定する必要がある。これは上記のDXの視点では、業務文書を学習データとして、業務の判断を支援するAI機能を学習・構成することに相当する（Fig. 2）。

2-3 言語モデルのトレンド

2007年に提案されたWord2Vec²⁾は、連続値ベクトルによる単語表現を導入し、言語モデルを機械学習により構築するというエポックメイキングな技術であった。しかし表現能力自体は低く、例えば同義語や未知語などの問題は解消できなかった。2018年に提案されたELMO⁴⁾及びBERT⁵⁾は、言語モデルの学習と、その言語モデルに基づくアプリケーションの学習を分離した事前学習—ファインチューニングフレームワーク（Pretraining-Finetuning framework）に特徴があり、大量のコーパスをもとにした自己教師付き学習（self-supervised learning）と周辺単語の分布を加味したエンベディングによりコンテキストを考慮した単語表現を獲得することで同義語の問題を解消している。また単語を部分に分解したうえで符号化するサブワードエンベディング（subword embedding）技術により、語彙として学習されていない未知語に対してもある程度の頑健性（Out-Of-Vocabulary耐性）を確保している。これら様々な工夫により、産業応用のユースケースで求められる実用的な性能を達成している。

これまでにBERTをもとにした様々な派生モデルが提案されている。これらの派生モデルを含めBERT世代の技術の多くはTransformerのエンコーダデコーダ構造のうち特にエンコーダに基づいているため、識別や回帰などの自然言語理解（Natural Language Understanding）に適しており、既に産業

応用として広く実用化されている。これに対して自然言語生成（Natural Language Generation）課題は、GPT-3などのデコーダに特化した大規模言語モデルが適しているが、自然な文章を生成するために、大規模な言語及び知識モデルを内包するの必要があり、学習、推論共に極めて大規模な計算機環境を必要とする。文書の生成は重要な産業応用であるため急速に技術開発が進んでいる。

2-4 自然言語処理アプリケーションのトレンド

前記の言語モデルのトレンドに合わせ、自然言語処理アプリケーション及びその開発にも傾向の変化が見られる。

第三次AIブーム以前は、機械翻訳などをはじめとした自然言語処理アプリケーションでは、目的ごとの言語モデルを開発し、且つ入力から出力までの流れを細かなステップに分割することで開発性と管理性を高めた、処理プロセスを分割統治するアプローチが主流であった。このアプローチでは特徴設計はアプリケーションごとに開発行為の一環として行われていた。これに対してBERT及びその派生モデルでは、多くのステップを1つのモデルに統合する形でアプリケーションを開発するアプローチ（End-To-Endアプローチ）をとっており、これによりアプリケーションごとの特徴設計の自動化を実現している。このアプローチに基づくモデルは、言語層、アプリケーション層それぞれの層数は変化するものの、モデル構成自体はEnd-to-Endアプローチとしてシンプルな構成になっている。このアプローチではさらに言語モデルを特定のドメインに適応させ、特徴設計を更新することで後段のアプリケーションの性能を高めるための転移学習によるドメイン適応が行われることがある⁶⁾ (Fig. 3)。これらの優れた特徴によりBERT以降のモデルは現実的な課題への適用性と高い精度を達成している。

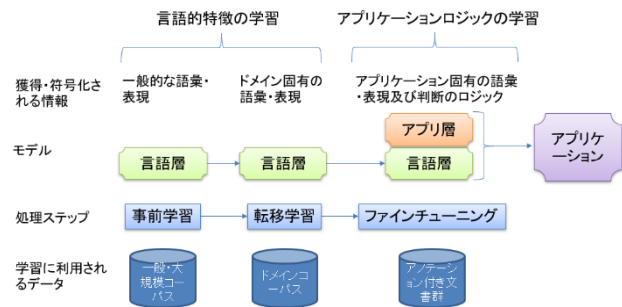


Fig. 3 BERT-generation model training procedure.

3. 「仕事のAI」

上記のAI、DX、言語モデル、自然言語処理アプリケーションの流れを踏まえ、リコーは個社の文書を解析する機能を開発し、2021年に仕事のAI事業を開始した。

本事業では、顧客企業の持つ文書データをリコーがAI機能に変換し、同顧客企業がその機能を利用する。中核となる技術は、実際の文書情報から、識別やスコアリングなどの自然言語理解の諸機能を学習し、個別サービスとして提供するための一連の機能群である。またこれらと並行してリコーが収集したデータに基づき学習・開発した一般向けサービスも展開している。これらのサービス群を効率的に開発・維持するために、MLOps機能、及び自然言語理解機能を開発するためのツール群を整備した。

3-1 提供機能

2022年時点で仕事のAIでは以下のお客様の声活用支援サービス群を一般向けサービスとして提供している (Fig. 4)。これらのサービスでは、お客様の声（Voice of Customer: VOC）を自動的に分類、定量評価する機能を提供している。



Fig. 4 仕事のAI (AI for your business) services.



Fig. 5 仕事のAI (AI for your business) Offering.

また、これらとは別に個社の要望に対応するためのメニューとして、文書作成支援、及び営業支援を設定している (Fig. 5)。これらのサービス群、すなわち市場情報の解析 (お客様の声解析支援)、社内文書の作成効率化 (文書作成支援)、及びこれらの結果を活用した営業活動の効率化 (営業支援) により、業務ワークフローの多くのステップを組織横断的にデジタル化することで、顧客企業のDXに貢献する。

各サービスでは、それぞれ様々な種類の文書を入力として想定し、業務上の高度な判断をAIにより支援する。このために、業務文書からAI機能の学習・推論のために必要な以下の機能群を開発した。

- (1) 業務文書から情報を抽出する文書画像解析、文書構造解析、及び情報抽出
- (2) 上記情報に基づき業務上の判断を識別やスコアリングなどの課題に落とし込むための要求開発及びAI機能設計技術
- (3) 識別、回帰、抽出などの典型的な自然言語処理機能をまとめ、多言語及び多言語モデルに対応したAI機能モジュール群

3-2 AI機能モジュールの特徴及び性能

仕事のAIをはじめとした自然言語処理を活用したアプリケーションには、いくつか代表的な機能が

Table 1 Representative NLP tasks.

カテゴリ	タスク	機能例
自然言語理解	識別	営業日報やvocなどの文意による識別
	スコアリング	営業日報やvocなどの定量評価
	抽出	営業日報やvocからの根拠文抽出
	系列対応	契約文書間の比較など
	クラスタリング	営業日報やvocなどの文意によるクラスタリング
自然言語生成	系列生成	営業日報やvocなどの要約など

ある。これらを社内で広く使えるAI機能モジュールとして整備し展開を進めている (Table 1)。これらのうち、自然言語理解についてはBERT世代の公開言語モデル及び技術をベースに開発し、他言語の言語モデルの入れ替え、及びRoBERTaやELECTRAなどBERT世代の他の言語モデルへの入れ替えを可能にするアーキテクチャにしている。

このAI機能モジュールが提供する機能のうち、識別、スコアリング、抽出、クラスタリングは、例えば文書種類の識別や、影響度の数値化、抽出型要約、及び文書群の構造化など、一般的な用途として広く用いられるものである。

3-2-1 背景技術

業務アプリケーションをBERT世代の言語モデルをもとに学習する場合、必要な教師データを縮減しつつ高い性能を実現するために、1) データ開発時の工夫、及び2) モデル学習時の工夫を行う。以下の説明では簡単のため対象機能を識別に限定して説明するが、他の機能についても同様の性能向上のための施策が取れる。

データ開発からモデル学習、デプロイまでを同じ仕組み・ルールの下で行うことで継続的インテグレーション及びデプロイ (Continuous Integration and Deploy) を行うMLOpsを本件に適用するため、本件のデータ開発における工夫とモデル学習における工夫について説明する。

3-2-2 データ開発時の工夫

現実的な応用では、実際の業務データの全てがモデル開発に使えることは少なく、与えられた業務データをキーワード検索などで絞り込み、無用なデータを省いたうえで、アノテーション作業に供することが多い (Fig. 6)。このキーワードを用いた検索が後段の学習に供するデータを定めるため、特定の顕著なキーワードのみに依存して学習用データが定まることのないようにキーワードの選定及び検索条件の設定を行う必要がある。キーワードの積集合で条件を定める場合、全てのキーワードが含まれる対象のみが学習に供されることになり、これらのキーワードをキューとした識別器が学習される可能性が高まる。逆にキーワードの和集合を条件とした場合には学習に利用できる対象は増えるがノイズも増すことになる。今回の開発では、データの偏りを緩和するために、キーワードによる全文検索だけでなくランダム抽出したデータもアノテーションの対象とし、ウェイトを調整しながら反復的に学習用データの抽出条件を設定している。

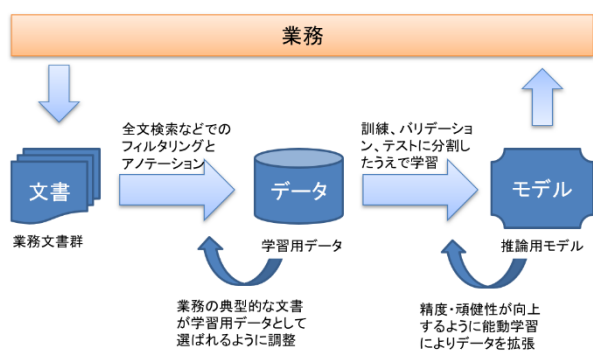


Fig. 6 Model development.

またアノテーションコストの低減のため能動学習を活用し、学習用データのうち判別超曲面に近い、もしくは判別能への寄与が大きい領域に対しては追加のデータが選ばれるように前記抽出条件を調整 (uncertainty sampling) している。併せてデータ拡張 (data augmentation) 技術も利用した学習データ

の追加を行うことでモデルの精度と頑健性を向上させる。

識別アプリケーションの学習には、データ開発時に識別カテゴリ自体が業務の要求を満たしつつ、重複などの無駄がなく構成されていることが求められる。また、アノテーションの付いていない業務データに対し、データ開発者 (アノテータ) が正確にラベルを付与するためには、カテゴリ自体が合理的に構成され、且つ各カテゴリに対してアノテーションに十分な情報が提供されることが必要である。そのため我々のアプローチでは、大カテゴリ、中カテゴリ、小カテゴリの順に詳細化し、実際の識別カテゴリは小カテゴリとして具体的に記述できるものとしている。事例をTable 2に示す。

Table 2 Label category example.

大カテゴリ	中カテゴリ	小カテゴリ	定義	キーワード
経営課題	事業継続性	人事系	後継者 跡継ぎ 後任者 ...	後継者、後継ぎ、事業継承、引継ぎ、人材不足、人材確保、...
		資金繰り	融資 資金不足 ...	資金繰、融資、資金不足、運転資金、財務体質、...
	
	事業変化	新規事業	事業展開 新規事業 ...	事業変化、新規事業、事業展開、事業拡大
		海外展開	海外拠点 現地法人 ...	海外展開、グローバル、海外拠点、海外法人、...
	

データアノテーションの粒度、及びシングル/マルチラベルの選択も識別アプリケーションの精度に大きな影響を与える。コンテキスト情報を拡張し精度を向上させるために、学習及び識別の両方で、なるべく多くの情報を入力として与えたいが、BERT世代のモデルでは入力トークン長に制約がある。入力文章を文に分解し、上記制約を考慮したうえで識別精度に影響を与える文を選択的に入力として用いるため今回は入力からのノイズ除去の後、利用する言語モデルが許す入力トークン長までの情報を与えている。またアノテーションを文単位で行うか、レコード単位で行うかも性能に大きな影響を与えるため、レコード単位、文単位、スパン単位でのアノテーションを使い分けている (Fig. 7)。

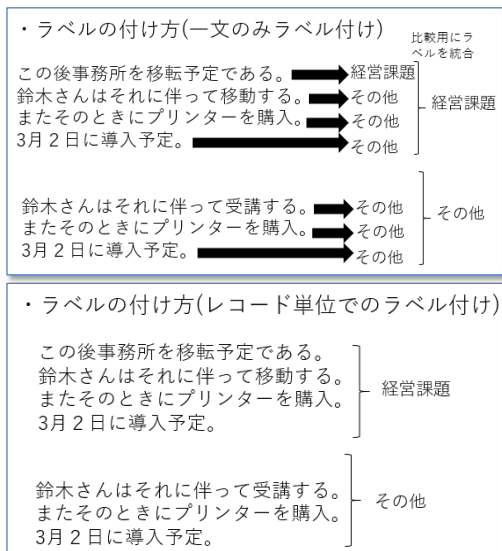


Fig. 7 Labeling example.

3-2-3 モデル学習時の工夫

機械学習を利用した識別や予測などの機能の精度が低いケースでは、良い特徴量が得られていないことが原因であることが多い。複数の異なるタスクで共通の言語モデルをファインチューニングすることで、タスク間で共通する頑健な言語表現(特徴)を獲得し、各タスクを高精度化するために、マルチタスク学習を利用している。これは疑似的なデータ拡張とみなすこともできる (Fig. 8)。

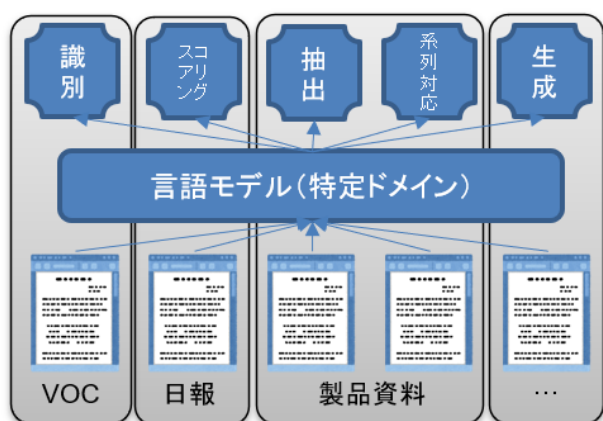


Fig. 8 Multi-task learning.

業務データを前提とする実際の応用事例では、識別クラス間でのインバランスが問題になることが多い。これは自然言語処理だけではなく、他のドメイン、例えば物体認識などでも広く見られる傾向であり、クラス間のデータ量不均衡を是正するためにオーバーサンプリング・アンダーサンプリングを組み合わせて適用することが多い。

例えば特定の事象を文書から抽出するある事例の場合、抽出対象は「その他」に対して約1,000分の1しか含まれておらず、単純にバランシングを行うと大半の「その他」の情報は無駄になってしまう。BERT世代の言語モデルを活用した自然言語処理に基づく業務支援アプリケーションの開発の場合、「その他」の文書も対象ドメインの情報を表しているため、表現学習のドメインアダプテーションに活用することができる。例えばFig. 7の例における「その他」に含まれる「働き方改革」や「プリンターを購入」等は識別したい経営課題ではないが、VOC中で顧客企業の興味の対象を表す「VOCらしい表現」である。エポックごとに上記のオーバーサンプリング・アンダーサンプリングの組み合わせを調整することで、このような情報をモデルに取り込むことができる。またこれは疑似的なアンサンブル学習とみなすこともできる (Fig. 9)。

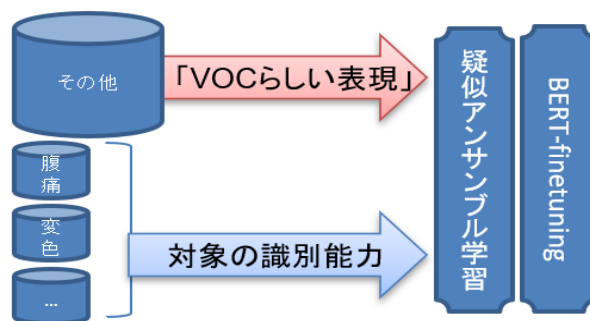


Fig. 9 Pseudo-ensemble learning.

4. 評価

前記の内容を含め、実際のアプリケーション開発では様々な工夫を案件ごとに行い、精度と頑健性を向上させたモデルを開発する。研究として精度評価と改良をシステムティックに行うためには、適用した工夫の影響を調べるアブレーションスタディ (ablation study) が必要だが、本稿では簡単のため基本技術との比較評価及びケーススタディをそれぞれ1例のみ示す。

4-1 BERTとの性能比較

本件では最初期にBERTをベースに開発を開始した。そのため以下にBERT (所謂Vanilla BERT) との比較結果を示す。実験設定は以下の通り。

モデル: bert-base

学習・試験データ: 当社の営業日報データ

タスク: 識別タスク (5値識別)

学習サンプル数: 5,000

ベースとなったBERTに比べ、Precision, Recall両面で優れた性能を持つことがわかる (Table 3)。

Table 3 Comparison result.

マクロ平均	BERT	当社モデル
Precision	0.650	0.852
Recall	0.155	0.591

4-2 リコーコールセンターでのケーススタディ

リコーはグループ内に複数のコールセンター機能を持ち、製品群ごとに管理している。これらのコールセンター機能を担う組織では多くのオペレータが入電を受け、結果をテキストとして記述し報告している。製品やサービスに対する問い合わせの内容、及び当社のサービスとしての規定に応じて、例えばお客様の問い合わせ・ニーズが特注対応できるケースや、対応できないケースそれぞれに対応方針が決まっている。現在は、事前の製品及び判断基準の学

習をしたオペレータが上記の判別を行っているが、これを自然言語処理で支援するために、前記の仕事のAI公開サービスを含め、複数の識別機能を適用した。

特定の2種類の商材について、仕事のAIサービスによりニーズの有無を判断した結果をTable 4及びFig. 10に示す。本件では予め試験用データとしてオペレータによる判別結果 (正解情報) を付与したデータを用いて評価を行っている。PrecisionとRecallはトレードオフの関係にあるが、今回はニーズとして抜け漏れ少なく後段のタスクで処理することを重視し、Recallをほぼ80%になるように調整した結果、Precisionは約65%となった。

Table 4 Evaluation results in 仕事のAI (AI for your business).

Recall	79.4%	463件のニーズのうち、368件を見落とさなかった
Precision	64.5%	570件をニーズと判断したうち、368件が正解だった

		人による分類(正解)		計
		ニーズ	その他	
AIによる分類	ニーズ	368	202	570
	その他	95	3,460	3,555
計		463	3,662	4,125

Fig. 10 Comparison in 仕事のAI (AI for your business) service.

さらにこのニーズの有無を大分類として、個別の識別モデルを開発した。分類内容をTable 5に示す。

この分類はコールセンターの業務をヒアリングし、AIによる判別結果の利用シーン (ユースケース) を想定して設定したものである。ニーズの有無による大分類から、契約のカバー範囲か否かの詳細に至るまでカテゴリを階層化することでAI機能の要求開発を多くのステークホルダに対して簡明なものとし、データアノテーションを行う際のアノテータへの作業指示もより明確になるようにしている。

Table 5 Labeling structure.

<分類>				
No.	大分類	中分類	小分類	内容
0	その他	その他	その他	下記いずれにも該当しない
1	ニーズあり	できないことあり	契約内	提供中の契約に関するお問い合わせのうち、契約対象内の内容
2	ニーズあり	できないことあり	契約外	提供中の契約に関するお問い合わせのうち、契約対象外の内容
3	ニーズあり	できないことあり	追加機能	追加の機能提供が必要な案件
4	ニーズあり	業務連携	業務連携	他組織との連携が必要な案件

上記構造化された小分類での分類性能をFig. 11及びFig. 12に示す。小分類単位での精度は5分類のマクロ平均を用いて計算している。下記性能に至るまでに1サイクルの評価及び教師データの作成・追加を行った。一般的なベンチマークタスク、データに比較して低い数値となっているが、これは現実の入電が様々な言い回しで記述されていることに依存する。

Fig. 12では、各クラスでの精度に大きな違いがあることがわかる。これは識別対象の各クラスがどれだけ特徴的なテキスト上での表現や語彙を持っているか、他のクラスと比較してそれらがどれだけ顕著であるかに依存する。

	precision	recall	f1-score	support	
分類番号	0	0.97	0.94	0.96	3663
	1	0.66	0.83	0.74	64
	2	0.51	0.73	0.60	147
	3	0.51	0.87	0.64	129
	4	0.79	0.41	0.54	122

Fig. 11 Evaluation result (fine labels).

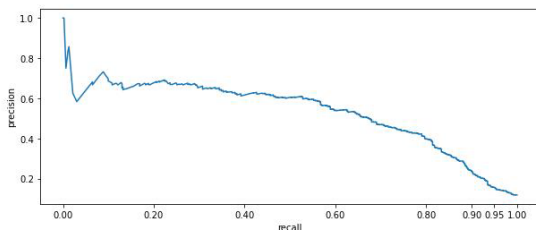


Fig. 12 PR-curve (fine labels).

5. まとめ

本稿では顧客企業のDXに貢献するための「仕事のAI」事業と技術的背景、及び本件の文脈におけるDXとAIの関係について説明した。

DXの2つの側面のうち組織横断のプロセス効率化を、AIをはじめとしたデジタル関連技術で促進するという考えに基づき、実際の業務で用いられる文書をもとに業務判断の支援機能を学習・構築するための頑健な自然言語処理システムを開発した。モデル学習のためのデータ開発と学習戦略を整合的に定めることで、実際の業務において実用的な精度を達成した。

参考文献

- 1) R. Salakhutdinov, G. E. Hinton: Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure (2007).
- 2) T. Mikolov, et. al.: Efficient Estimation of Word Representations in Vector Space (2013).
- 3) デジタルトランスフォーメーションの加速に向けた研究会 中間とりまとめ (参照2022-12-10), https://www.meti.go.jp/shingikai/mono_info_service/digital_transformation_kasoku/20201228_report.html
- 4) M. E. Peters, et. al.: Deep contextualized word representations (2018).
- 5) J. Devlin, et. al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).
- 6) S. Gururangan, et. al.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks (2020).