札幌 AI ラボ　AI 普及啓発セミナー
「DX を支える自然言語処理」
～事例から見える社会との親和性とビジネスの種～　実施報告（抄）

開催日：2022 年 8 月 24 日（水）15：00～17：00
場　所：札幌市産業振興センターセミナルーム D＋Youtube ライブによるオンライン配信
共　催：一般財団法人さっぽろ産業振興財団
共　催：札幌 AI ラボ
参加者：37 名（会場 13、配信視聴 24）
プログラムと内容概略（以下、敬称略）

1**【講演】山下氏による講演**

北海道大学大学院情報科学研究院 情報理工学部門 複合情報工学分野
調和系工学研究室 准教授
札幌 AI ラボ　テクニカルメンバー 山下　倫央

●DX を支える自然言語処理
・調和系研究室の紹介
学術的に価値の高い研究している。
人の感性を理解するなどの、個人レベルの支援は数多くあったが、
社会の円滑化を考えることができる AI 集団としてもうまく回るように研究をすることをコンセプトとしている。

●自然言語処理の基礎
・AI システムの市場規模
国内 AI システム市場が 2026 年までに 3500 億円から、8000 億の規模に拡大予測されている。

日本自然言語処理（NLP）市場も 2030 年末までに 670 億から 6000 億の収益を獲得すると予測されている。

・自然言語処理技術のサービス利用事例
言語は方言などの地域差があり、構造も複雑だが、人間が日常的に利用している自然言語を計算機で扱うことができれば、蓄積した膨大なテキストデータの有効利用が可能になる。

自然言語処理技術のサービス利用例
機械翻訳、検索エンジン、感情認識、テキストマイニング、AI スピーカーなど

製品レベルの利用例
AI　EXPO2022 春で目立っていたブースは、チャットボット。
感情認識は、数多くの分野に関連する自然言語であり、クライアント対応や社内コミュニケーションで人気があった。

・近年の自然言語処理の発展要因
スラックなどのコミュニケーションツールが普及し、データを蓄積し、人手では困難な膨大なテキストデータの解析ができるようになった。
議事録の自動生成や、押印の省略を目的とした紙媒体の減少など。

高性能の言語モデルの開発
マスク化言語モデル　BEAT　（Google）深層学習モデル
文章生成言語モデル　GPT-3　（OpenAI）

文字認識技術の向上から、テキストデータ抽出のレベルも向上

・自然言語処理の概要
自然言語とは、日本語や英語などの人間が普段使っている言葉。
自然言語処理とは、言葉をコンピュータに理解させるための技術分野。
5 段階の処理が行われる。（前処理、形態素解析、構文解析、意味解析、文脈解析）

前処理
そのままでは処理しづらいテキストデータを整備
句読点の統一（。、）や、「えー、あー、こんにちは」などの言い澱みを分析可能な形に置換

形態素（けいたいそ）解析
文章を形態素に分割して、意味や品詞などを判別する。
例：　メロスは激怒した　→　メロス　　は　　激怒　　し　　た

構文解析
形態素解析で取得した単語間の関係性を解析

メロスはセリヌンティウスと羊を食べた
メロスが「セリヌンティウスと羊」を食べた

意味解析
メロスはセリヌンティウスと羊を食べたは２つの解釈ができる。
① メロスが「セリヌンティウスと羊」を食べた
② 「メロスとセリヌンティウス」が羊を食べた

メロスとセリヌンティウスは人名であり、羊は一般名詞であることから、②を導く。

文脈解析

複数の文について形態素解析と意味解析を実施し、文同士の関係性を解析する。
こちらは研究レベルの技術なので発展が望まれる。

意味解析と文脈解析は自動化するために数値化して、ベクトル空間に埋め込む。

・最新技術の紹介
文章生成 AI「GPT-3」が Reddit で 1 週間誰にも気付かれずに人間と会話していたことが判明した。
単語を次々に予測できて、回答のスピードがとても早い。
AI 俳句もこれを推奨

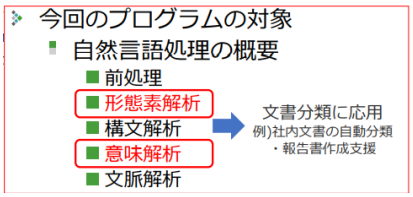
「DALL-E」
文章から画像の生成ができる。
an illustration of a baby daikon radish in a tutu walking a dog.
チュチュを履いた赤ちゃん大根が犬と散歩している絵がたくさん出てくる。

●調和系工学研究室の研究紹介
・AI 一茶くんの開発
人間が読んだ俳句の言葉のつながりを学習して、俳句を 1 分間に 400 句のペースで昼夜問わず生成し続ける。人間が詠むような心を動かす俳句を見分けることは難しいが、音数や季語など、決まったルールにあった処理を得意とする。

・人工知能による競輪予想記事の自動生成

・帝国議会議事速記録
テキストデータの構造化
新しい単語の検索　準備を進めている。

●まとめ
～Sapporo AI Lab AI 人材育成プログラム～ 2022 年度版
「実践的データ分析講座（テキスト分類）」（1 セット目：9/8&9/15 開催）を受講することで、形態素解析や意味解析に係る部分の技術的習得が期待できるので、受講をお勧めしたい。



・DX に向けた自然言語処理の導入ポイント
蓄積した社内文書の利活用
記録されている媒体で、分析の労力が異なる。
そのデータに分析する価値があるかどうかの判断が必要。

単一目的では高コストであり、多目的な利用を想定すること。
これからユーザーから集めるデータは、人の手を減らすようにすることが重要。
自然言語処理はあくまでも手段であり、デジタイゼーションや蓄積データの利用ではなく、新たな価値を生み出すデジタライゼーションを目指す。

2　山下氏と石田氏の対談

株式会社テクノフェイス
代表取締役
札幌 AI ラボ　テクニカルメンバー　石田　崇

石田氏
AI ラボ札幌市コールセンター実証実験を実施した。
５年前、チャットボットはあまり普及していなかった。
大量にデータを持っているものは何？と考えたときに、「札幌市の電話の応対記録だ。」となり、やってみることになった。
コールセンターの方達の努力もあり、150 万件の応答記録が綺麗なデータで残っていたため、前処理がてにをは含めて、想定していたよりも大変ではなかった。
通常はシナリオ型でツリー上のものがきれいにできるのだが、１５０万件のデータを入れて、さあ答えてくださいの状態でも、７，８割くらいのものが意図した FAQ にたどり着く成果を見た。

山下氏
自然言語処理はこの５年で GPT-3 など色んなモデルが出てきて進化した。
前処理が済んでいたら、モデルを変えて精度の向上が図れるのではないかと思うが、取り組んでいることはあるか。

石田氏
技術進歩を受けて、その辺りの興味はあるが、個人情報など、てにをはの前処理だけじゃない取り扱いは明確にしないといけないところがある。

山下氏
プライバシーの問題があるなら、前処理の前の処理をしてデータを共有可能な状態にするのがいい。

石田氏
文章としてはしっかりしているが、１５０万件のデータは多岐にわたる。
オープンデータ化するのなら良いデータだと思うが、行政でそれを行うのは容易ではないというのは聞いたことがある。

山下氏
業務改善の観点からもいいデータだと思うが、広報活動などに活用できるといいと思う。
100%の質問に答えるよりも、質問が来なくなることで、「札幌市いいね」と言われる DX の第一歩だと感じた。

●前処理の話があったが、相談する案件の中で、今はどんな感じ？

石田氏
計画書を作る業務が重たい。
毎日数十枚だとしんどい。同業者でも大変な業務になっている。
チェック項目をデジタル化する。できれば数値化して扱いたい。ベクトルで書けるのか。
データの単純化を考える。
データ数が少ないのであれば、まずは自然言語を使わないでデジタルでできないか考える。
残りのところを部分的に自然言語処理を使う。
機械学習に本当に必要なデータがわかるようになる。

文書生成は今までの報告書のどれだけパターン化できるのかを解析する。
AI でやりたい依頼を受けて、システムになる割合も多い。
対話するように会話を出す仕組みは研究レベルではどうなっているのか。

山下氏
この文章不自然だよねとか、いい俳句ってなんだっていう評価が難しいところもある。
目的が大事で報告書を受ける人が何を知りたいのか定義した方が世の中のためになる。

●文書生成について、実際に直球の質問がきたときにどうする？
山下氏
精度を指標にすると、正確な分類は人間でも難しい。目標にはせず、半分の質問を答えてくれたら、半分の時間が浮くと考えたほうが良い。
優先順位の高いところから潰していきましょうのモチベーションになる。

石田氏
チャットボットは問い合わせが多いパスワードの発行を自動発行にとてもいい。それに追われている人は少し業務が楽になる。文脈を読むのが AI、解決するのがシステムというのが真骨頂ではないかと思っている。

●市場拡大予測があるが、研究、解析などで気になるところはあるか。
山下氏
文章のクオリティも重要だと思うが何を求めているのか、受け取る側の求めるものに沿うようにする。
細かいところの DX を念頭におかなくてもいいのでは？
FAX を OCR に変えるのではなく、メール添付にしたらどうか、など。
対応できる人は手間もコストもかからないようにする基本方針が大事ではないか。

石田氏
FAX で「いつもの」と手書きで書いてくるお客様がいる中で、AI で文脈を理解しようとするのが難しい。手間や分析に時間がかかる。iPad の使い方を覚えるなど、人間が機会に歩み寄った方が DX が進む場合もある。

●ジャンルは多種多様だと思うが、次のマーケットはどのような領域があると思うか？

山下氏
テンプレートの判断は出てきてもおかしくないと思う。文章の候補をそれなりの成分を作るといいものが出てくるのではないか。
AI 俳句でも使っているが、人間がやると大変なブレインストーミングで使う。

●帝国議会議事速記録の研究の意図はなにか。
山下氏
企業から報告書があるが、いい使い方がないか？や、ZOOM 会議で議事録が蓄積されていいことがあるか？と聞かれることがある。
Todo を書き残しておいて忘れないようにするのが重要。
どんな文章を解析するのが大事。
現段階で国のレベルで最新の情報で、ボリュームがあるのは分析のやりがいがあるというところを考えると、後付けではあるが議事録だということになった。

3　質疑応答及び今後の予定

質問二件
A
医者の先生に質問した時に、パーソナライズされた返答が返ってくるチャットボットができる可能性があるのか。また、それはいつごろか。

山下氏
デジタルツインは狙われているところではあるが、医学は知見が膨大であり、非言語的なものが多いので簡単ではない。簡単な診断はできるようになるかもしれない。今まで見つけたことがない初めての病気を見つけるのは難しいかもしれない。

A
AI 診断は東大で進んでいるという話は聞いたことがある。人の命関わってくるので、責任がある。大御所の先生によって判断が違うことがある。責任というところで判断の参考になるものがあればいいと思った。

石田氏
怖い先生のデジタルツインを作るということであれば、その知見を引き出さないといけないので、質問しない為に、たくさん質問しないといけないですね。

B
チャットボットについて
チャットボットを使うと満足度が下がってしまうと思う。
寄り添う AI が大事ではないか？いい技術はあるか。

山下氏
コミュニケーションにおいてデータが少ないというのがよく聞く。
理解して共感して問題解決をするのが一番いいと思うが、まだググった？というくらいカジュアルに検索できる領域に達していない。
わからない話を助ける AI がいるか、研究者がそのスキルを身に着けるの二択。

B
音声認識の方言について、前処理の課題は？

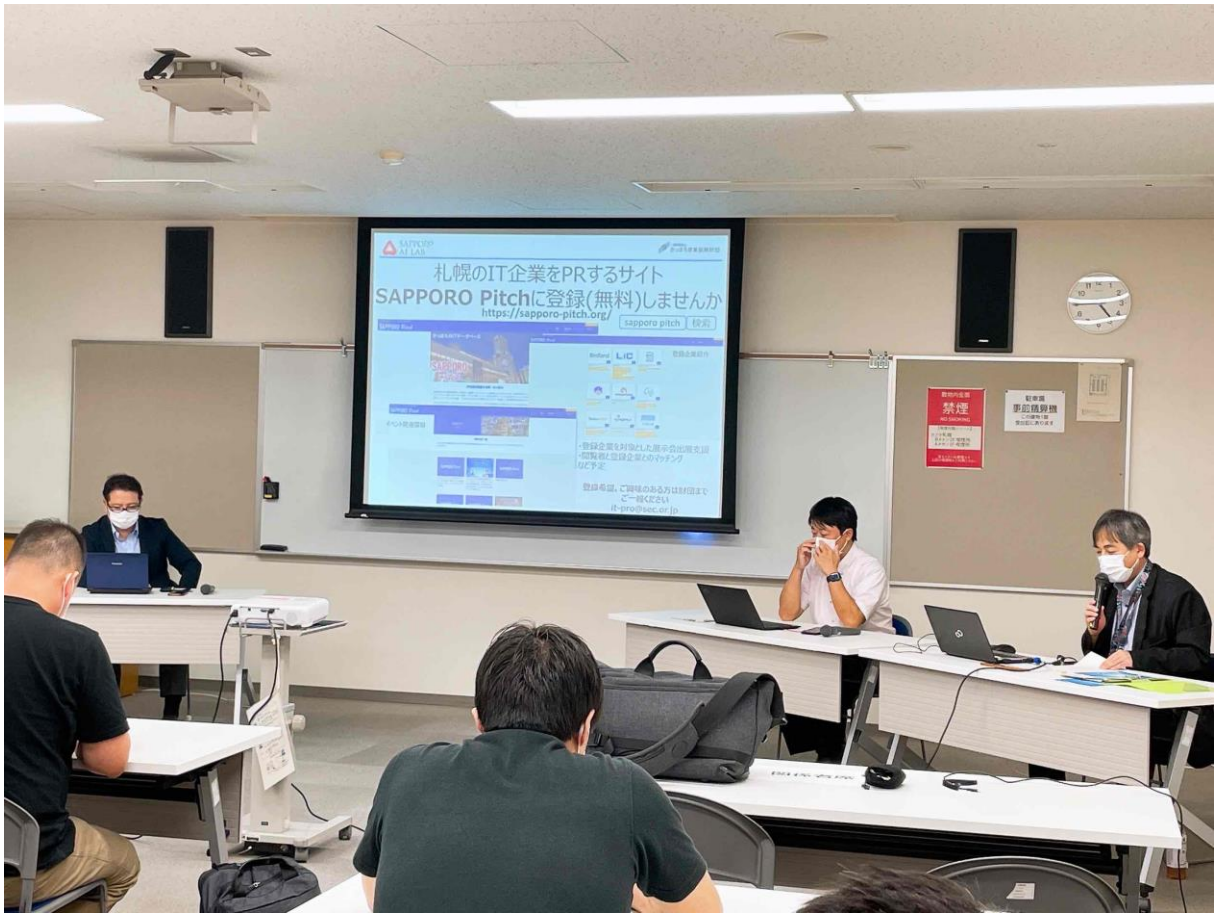
山下氏
音声はまだまだの部分がある。テキストに落とせたら、方言はなんとかなると思う。英語や日本語でもできるのであれば、青森弁でもできると思うがデータの数問題だと思う。

石田氏
認識する際に辞書との突合せが発生するのである程度正しい日本語である必要がある。辞書から作り直さなければいけないという話になると思った。

B
一昔前に自然言語処理を使って大学受験を頑張りましょうというプロジェクトがあったが、今の自然言語処理の技術を使ったらどうなのか。

山下氏
一気にあがらないのではないかということで、今は終了している。偏差値は５７くらいまであがった。外部の知識をどれだけいれるかが難しい。それだけのコストをかけて偏差値をあげるのかということに価値を見出せない。
AI をどういう風に使うかの問題であり、AI が検出できない問題を検出するのはありだと思う。問題の中で完結する問題なのか、外部の知識を必要とする問題なのかでグレード分けするのが大事だと思う。

事務局からの連絡



◇アンケート協力依頼
◇データサイエンティスト
～Sapporo AI Lab AI 人材育成プログラム～ 2022 年度版
「実践的データ分析講座（テキスト分類）」（1 セット目）
１日目　テキスト分類の要素技術について（座学と演習）
令和 4 年（2022 年）9 月 8 日（木）14:00～17:00

２日目　課題の報告と検証（座学と演習）
令和４年（2022 年）９月 15 日（水）14:00～17:00
（2 日間トータルで１つの研修）
2 セット目の１セット目と同内容。開催日は 11 月 22 日(火)＆11 月 29 日(火)

◇札幌の IT 企業を PR するサイト Sapporo Pitch の紹介
市内 IT 企業のデータベースとして運用中。登録は無料。登録を希望される方は、さっぽろ産業振興財団　山下（it-pro@sec.or.jp）までご一報を。

◇DX モデル創出補助金【二次(追加)公募】
公募期間:令和４年(2022)年８月１日（月）から９月２日（金）17:00 まで
事業対象期間：採択日（令和４年(2022 年)9 月下旬頃）から令和５年（2023 年）２月 28 日（火）まで
公募締切：９月２日（金）17:00 まで

以上