機械翻訳の応用分野：SQL生成技術の紹介

数理システムユーザーコンファレンス2018

㈱リクルートテクノロジーズ
データテクノロジーラボ部
牧 允皓

2018年11月22日

DATA TECHNOLOGY LAB

RECRUIT リクルートテクノロジーズ

自己紹介

氏名

牧 允皓 (まき よしひろ)

略歴

新卒でソーシャルゲームの会社に入社。
データサイエンティストとして4年間勤務。アクセスログの分析、施策の効果検証、異常検知システムの構築などを経験。
2017年にリクルートテクノロジーズに入社し、機械学習のソリューションを開発、運用するグループに所属。主な業務はA3RTのプロダクト開発・運用と、外部の企業との協業など。

学歴

九州工業大学大学院 情報工学府

その他

データサイエンティスト養成読本 登竜門編 共同執筆
機械学習の講師として活動

概要

1. リクルートのビジネス
2. データテクノロジーラボ部の役割
3. 今回のトピック：SQL生成

リクルートのビジネス

リクルートグループについて

創業	1960年3月31日　「大学新聞広告社」としてスタート
グループ 従業員数	40,152名　（2018年3月31日時点）
グループ 関連企業数	361社　（連結対象子会社、2018年3月31日時点）
連結売上高	21,733億円　（2017年4月1日～2018年3月31日）
連結経常利益	1,917億円　（2017年4月1日～2018年3月31日）
目指す世界観	 「あなた」を支える存在でありたい

RECRUIT
リクルートテクノロジーズ

リクルートの事業内容について

ライフイベント領域

ライフスタイル領域

旅行

じゃらん

エイビーロード AB-ROAD

グルメ・美容

ホットペッパー HOT PEPPER グルメ PRODUCED BY RECRUIT

HOT PEPPER Beauty

ブッキング テーブル

生活/地域情報

ポンパレモール

ケイコとマナブ

人間ドックのここカラダ

ビジネス支援

AirREGI

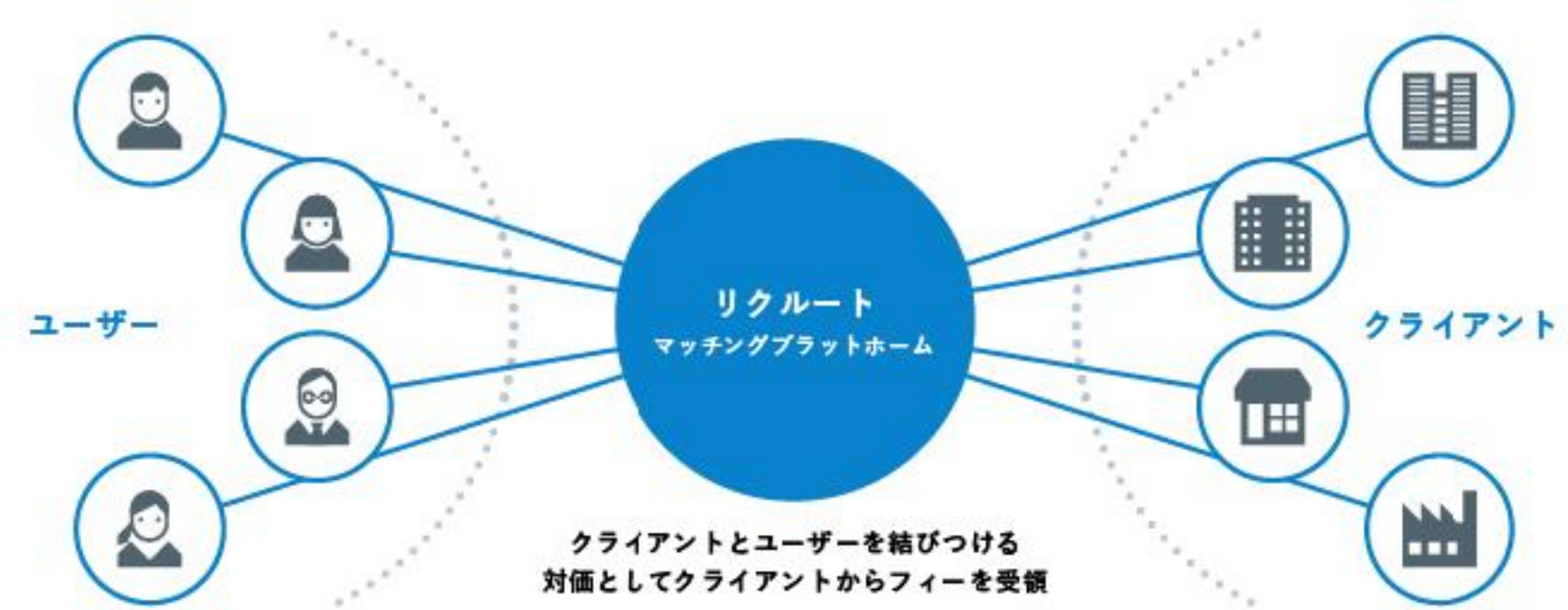
SALON BOARD

選択・意思決定を支援する情報サービスを提供し、
「まだ、ここにない、出会い。」を実現する。

RECRUIT
リクルート テクノロジーズ

リクルートのビジネスモデルについて

**リクルートには、ユーザーとクライアントという2つのお客様が存在します。
企業と人（B to C）、企業と企業（B to B）、人と人（C to C）、すべての間に立ち、双方にとって最適なマッチングを図る「場」を提供しています。**



**ユーザーとクライアントを新しい接点で結び、
「まだ、ここにない、出会い。」の場を創造する。**

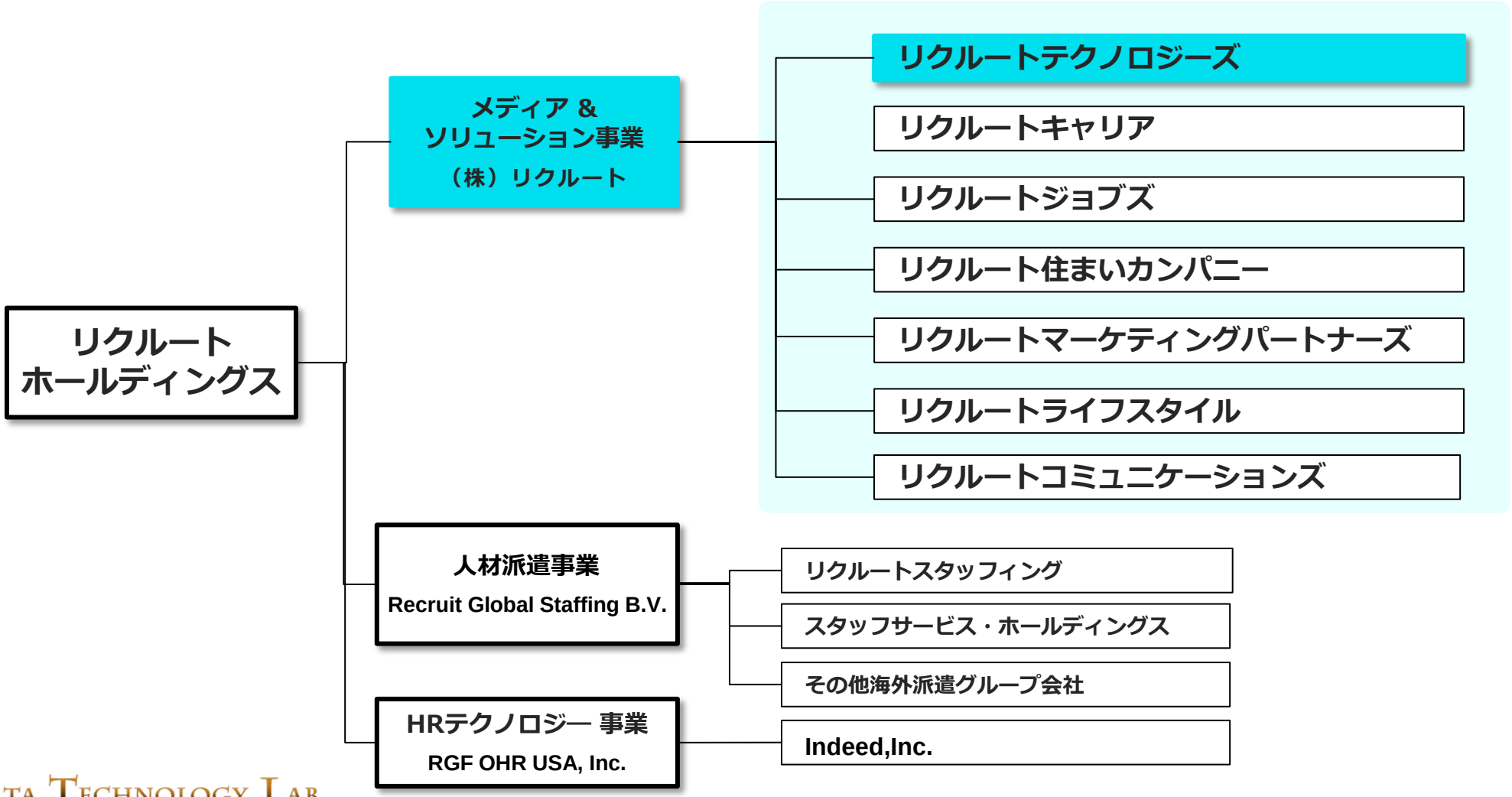
RECRUIT
リクルート テクノロジーズ

データテクノロジーラボ部の役割

DATA TECHNOLOGY LAB

リクルートグループにおけるリクルートテクノロジーズについて

リクルートテクノロジーズは、リクルートグループのIT・ネットマーケティング領域のテクノロジー開発を担う会社です。



DATA TECHNOLOGY LAB

リクルートテクノロジーズの役割について

技術・ソリューションを磨き続け、リクルートの各サービスがもつ価値を最大限に発揮できるようビジネスへ実装。

ITの側面からサービスを進化させることを通じて、世の中に新しい価値を提供していきます。

DATA TECHNOLOGY LAB

リクルートテクノロジーズの事業内容について

**将来のニーズを見据え、新しい技術のR&D・ソリューションの開拓を実現。
検証を続け、いち早く活用できるレベルに引きあげることで、中長期的なビジネス競争優位を構築していきます。**

リクルートテクノロジーズが担っているもの

DATA TECHNOLOGY LAB

データテクノロジーラボ部について

ビッグデータ、データテクノロジーを専門にする部署で、人工知能、機械学習と呼ばれる技術の研究開発をミッションにしている組織

リクルートテクノロジーズが担っているもの

データテクノロジーラボ部が担っている分野

DATA TECHNOLOGY LAB

データテクノロジーラボ部の取り組みについて：A3RT

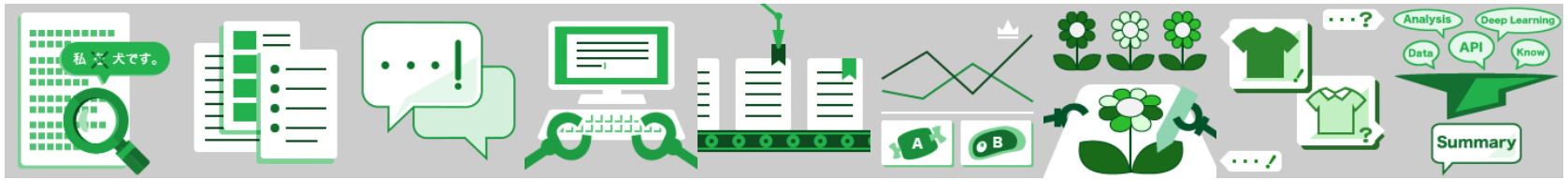
A3RTとは

- リクルートテクノロジーズが提供する機械学習のAPIサービス群
- リクルートグループが提供するサービスの価値を高めるために開発された
- 数年後に直面するであろうビジネス課題を想定し、最先端の技術の研究開発に取り組んでいる

プロダクト例

- 自動校閲：誤字脱字、誤表記などの文章校閲
- 文章分類：投稿された記事が規約に違反していないか判定
- 文章生成：原稿の自動生成
- 自然言語系以外に画像系のプロダクトも多数

公開されているA3RTの紹介



https://a3rt.recruit-tech.co.jp/

無料公開

- 2017年３月公開
- 内部のサービスに限定せず、様々なシステムに組み込まれることを期待

目的

- 多様なフィードバック
- モデルのブラッシュアップ
- 新しい使い方の発掘

今回のトピック：SQL生成

ビッグデータとSQL

ビッグデータの流行

- Internetの普及やストレージの低廉化などに伴い、ビッグデータという考え方が広まった
- ビジネスにおいて様々な場面でデータに基づく意思決定が求められるようになった

データ活用によって生まれた業務

- データを活用するために生まれた「データ抽出」、「データ集計」という業務
- 例えばデータベースに蓄積されたデータを抽出するにはSQLの理解が必須
- エンジニアやデータに係る技術者に集計依頼が発生

ビジネス課題

■ 顕在化しにくい集計工数

- 専門知識が必要であるにも関わらず、集計の工数は軽視される傾向
- 様々な組織でちょっとした集計業務が徐々に増加している(はず)
- 集計結果をみると別の新しい切り口で集計したくなるケースが多い

そこで、データ集計技術の大衆化を目指す研究を調査

Seq2SQL と WikiSQL

Salesforce Inc.

- 2017年に Seq2SQL に関する論文を発表
- Github上でデータセットが公開された
  https://github.com/salesforce/WikiSQL

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...	...	...	...	...

Question:
How many CFL teams are from York College?

SQL:
```
SELECT COUNT CFL Team FROM
CFLDraft WHERE College = "York"
```

Result:
2

Victor Zhong, Caiming Xiong, and Richard Socher. "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning." arXiv, cs.CL 1709.00103 (2017).

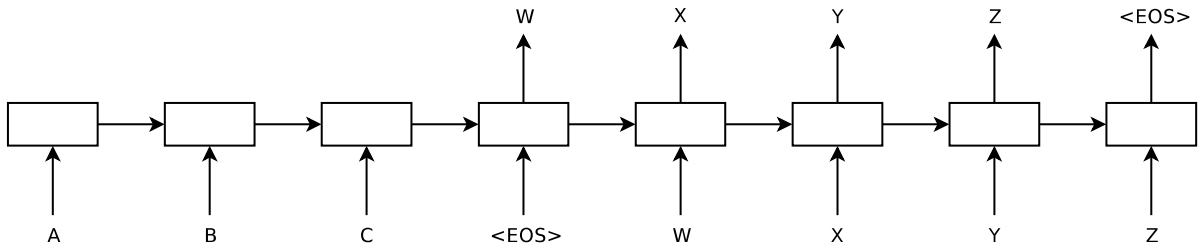
Seq2SQL が目指すもの

- Question から SQL に変換
- 未知のテーブル定義にも対応できる汎用モデルを構築することが目的
- 公開されたデータセットには幅広いテーブルに対して数組の Sequence と SQL を含んでいる

機械翻訳 Encoder-Decoderモデルについて

背景

- Sequence to Sequenceという考え方が2014年の論文で発表され多くの研究テーマに応用された
- Encoder-Decoder翻訳モデルともよばれるRNNから派生したモデル
- 以下のように”ABC”と入力すると”WXYZ”を出力する



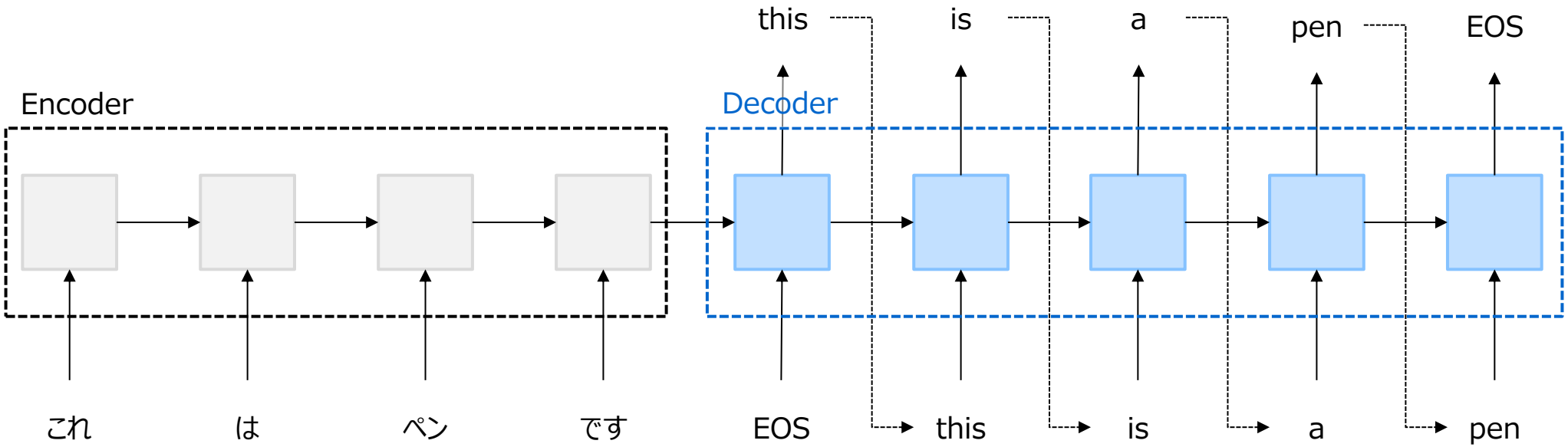
Ilya Sutskever, Oriol Vinyals, and Quoc V.LE. “Sequence to Sequence Learning with Neural Networks.” Advances in neural information processing systems. pp.3104-3112 (2014).

用途

- 自然言語の翻訳(日本語⇄英語)が代表的
- 今日では文章要約なども盛んに研究されている

Sequence-to-Sequence のイメージ図

単純な翻訳の例(これはペンです → this is a pen)



SQL生成のビジネス・インパクト

業務に潜むデータ集計

- ビッグデータ流行のピークが過ぎてもなお、多くの意思決定はデータに基づく
- 組織が大きいほど組織長が経営状態を把握するために集計業務が発生

潜在的な価値

- 蓄積された膨大なデータはDBで管理され、SQLを書いて集計する
- 組織長がデータサイエンティスト、エンジニアなどに集計を依頼する
- SQLを習得していない組織長が簡単にDBにアクセスできるとこの業務は減る

SQL生成による構造変化

従来

目指す世界観

DATA TECHNOLOGY LAB

問題の定式化

Seq2SQLのアカデミックなタスク

- 汎用モデルの構築(未知のテーブル定義に対してもSQLを生成できる)
- ビジネスの観点からは研究がまだまだ発展途上(精度が実用に耐えうるか不明)

解きたい問題を定義

- 汎用性よりも**高い予測精度**と**学習データの準備コスト**を優先
- テーブルが所与の状況で以上２点が現実的に実現可能か検証

学習用データの生成

データテーブルから学習データの生成

- Sequence と SQL を入力、出力文章として sequence-to-sequence のモデルを学習
- SQLの難易度が高い命令(JOIN や GROUP BY など)は初期段階では回避
- Sequence のバリエーションが十分になるようデータを準備する
  → かなり泥臭い作業で効率は悪い

学習データの自動生成

- 研究中のタスク
- 学習データ自体をテーブル定義から生成する仕組みを開発
- モデルの学習に十分なバリエーションかを検証中

開発(直近1年)

10-12月

・リサーチ開始

・古典的な Seq2Seq を開発

・テーブル定義を所与としてデモモデルの開発

・音声から制御できるUIの開発

1-3月

・Seq2SQL×音声UIの結合

・データ生成のスクリプトを開発

・β版完成

4-6月

・ラズパイ×Vioce Kitで実装

https://www.raspberrypi.org/

https://aiyprojects.withgoogle.com/

・無料公開の準備

7-9月

・8/23に無料版公開

・事業の業務効率化へ向けて

トライアンドエラーを繰り返しながらスピーディーに開発することで
フィードバックを得たり需要がある組織をヒアリングできた

DATA TECHNOLOGY LAB

A3RT Public API公開

無料公開

- デフォルトモデルで天気のテーブルに対してSQLを生成します

SQL Suggest API

DETAIL | NOTES | REFERENCE | SAMPLE REQUEST

SQL Suggest APIは、日本語の質問文をSQLに変換するプロダクトです。
SQLが書けなくても、質問文を入力すればデータベースになげるクエリを得ることができます。

https://a3rt.recruit-tech.co.jp/product/SqlSuggestAPI/

今後の展望

集計業務の自動化

- Google Home や Amazon Echo といったスマートスピーカーの普及とともに音声コマンドが浸透
- 部下に任せていた集計が一声で完了する世界観が実現可能

ビジネスインパクト

- 業務効率化を目的として導入を進めている
- Web画面、スマートスピーカーなど案件に最適な UI で提供できる

Google amazon

A3RT

どしどしご利用下さい！

https://a3rt.recruit-tech.co.jp/

RECRUIT
リクルート テクノロジーズ

メンバー募集中

DATA TECHNOLOGY LAB