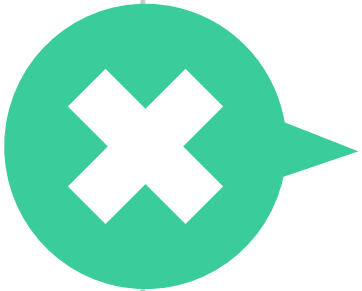
AI議事録の開発で感じたユーザー目線のAIの精度を測る難しさ

自己紹介

髙山　雄貴

- Nishika株式会社 リードAIエンジニア
  - 自社のAI議事録プロダクトのAIインフラの運用
  - 企業向けAIソリューションの開発/PoC案件のPM
  - AIの研究開発
- Kaggle Expert
- Cursorを使っているが、社内がClaude Code1択になりつつあるので乗り換え予定
- マイブームは半年前に始めたボルダリング



皆さんはどのAIチャットの
サービスをよく使いますか？

perplexity

Claude

AIの精度の良さを定量評価するのは難しい

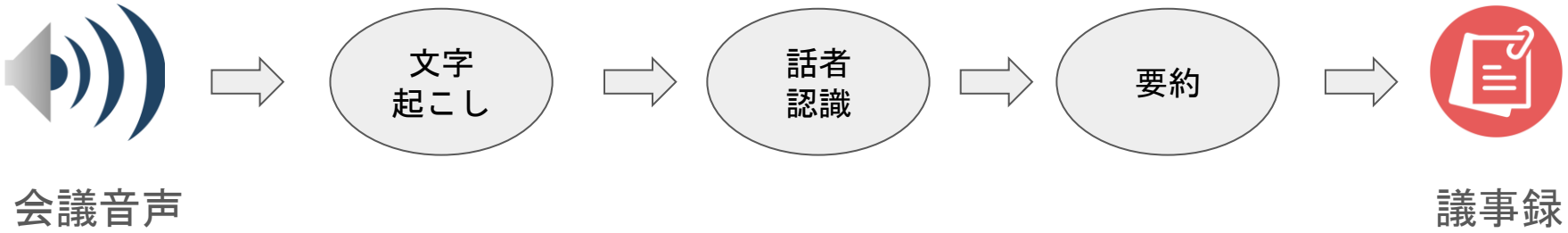
- 皆さん、どのAIサービスを使っていますか？→「精度がいいから」
- でも「精度」を言語化するのは難しい
- 測る物差しが多種多様
  - 例：コーディングの質問にどれだけ適切に回答できるか
  - 例：曖昧に質問しても、背景や意図を補ってよしなに回答できるか

AI議事録サービスのAI開発でも、ユーザーにとっての精度を測ることは難しい

SecureMemo

会議音声をもとに、AIが文字起こし・話者認識・議事録作成を一気通貫で行い会議の情報を資産化し有効活用できるAI議事録サービス

AI議事録を作成するAIのパイプラインは複雑



ユーザー目線の文字起こし精度は、CER（文字単位の誤り率）と一致しない

正解の文字起こし

「アサヒ商事との案件については予算は、予算上限を超過する以上、見直しが必要です」

不正解の文字起こし

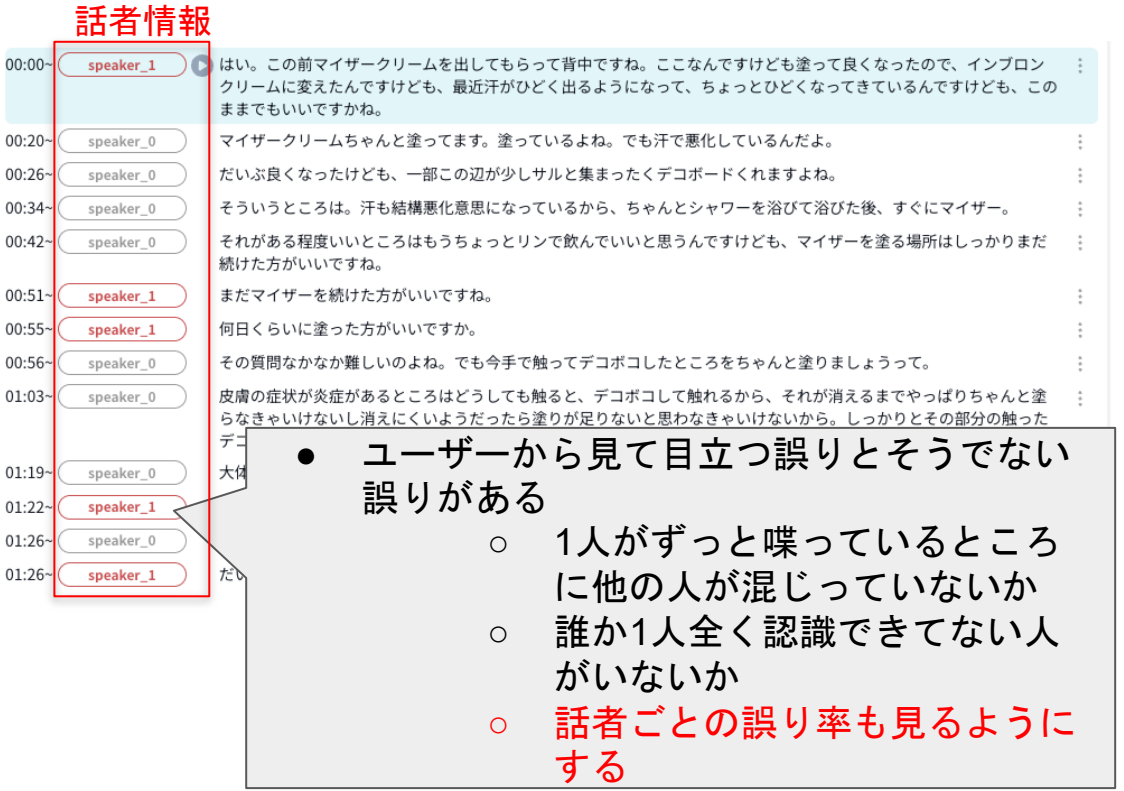
「**えーっと**、**アジア商事**との案件については予算は、予算上限を超過する以上、見直しが必要です」

- 大事な情報である会社名（固有名詞）を誤り、スコア以上に全体的に誤っている印象を持ちやすい
  - CER/WERと合わせて名詞誤り率も見るようにする
- 「えーっと」というフィラー（繋ぎ言葉）がが含まれる
  - 含めてほしいユーザーとそうでないユーザーで意見が分かれる

ユーザー目線の話者認識精度は、DER/IER（話者誤り率）と一致しない

DER/IERは

「誰がいつ話していたか」の正解と、AIの認識結果を時間軸で比較して、**話者がズレていた時間の割合を示す指標**



ユーザー目線の要約精度は、
内容・形式がユーザー自身の理想に近いか

- **議事録の要約精度は最も評価方法が悩ましい**
- 弊社ではAzure OpenAI Serviceを採用
- 内容面のゆれ
  - 人によって、重要視している会議の議論の内容のポイントは様々
- 形式面のゆれ
  - ユーザーによって求める議事録の形式は様々
    - 文字起こしの情報をあまり落とさない逐語録
    - 決定事項やTODOのみの箇条書き
- 弊社での運用
  - 評価観点を決めて、LLM as a Judgeと呼ばれるLLMによる定量評価の手法も導入しているが、見切れない観点もある
  - 人の目で見て、定性的に評価することも大事にしてモデルの選定を行っている

SecureMemoCloud 商談 _議事録

日時
- 2024 年 10 月 25 日 13:25

参加者
- 松田、鈴木、山本

決定事項
プランの決定
- チームプランを採用することを決定した。
  - 25 時間の利用時間を含む。
  - 要約機能を使用する。
  - 従量課金は適用しない。

契約形態の決定
- 年間契約を基本とするが、月契約の見積もりも依頼することを決定した。

議事要旨
利用時間とアカウント数の確認
- 録音を開始しました。ありがとうございます。早速ですが、本題に入ります。何時間利用するか、何人で利用するかの希望を伺い、それに基づいたカスタムプランを考えたいと思います。（松田）
  - 最大で実際は 9 時間程度ですが、バッファーを見て 1 ヶ月で 12 時間程度になります。（山本）
  - 12 時間ということで了解しました。（松田）
- 実際の文字起こしした結果をウェブ画面上で共有したい人数、アカウント数のイメージは何人でしょうか。（松田）
  - 台数は一台で良いですが、アカウント数によって料金は異なりますか？（山本）
  - 例えばチームプランのアカウント数は 10 名となっておりまして、ここまでは固定課金ですが、10 名以上増やす場合は従量課金となります。（松田）
  - 承知しました。であれば 1 アカウントで良いです。（山本）
  - 承知しました。（松田）

利用時間の再確認とプランの選定
- すいません。もう 1 回会議の時間を計算し直したら、7 月は 18 時間とかになるんですけど。（鈴木）
  - わかりました。（松田）
  - 25 時間でいいんじゃないですか。（山本）
  - そうですね。25 時間で本当に重い時だとそれぐらいかかってしまうので、ない時はないんですけど、ある時は 18 時間だなと思ったんですけど、いろんな会議があるなと思って。だったら 25 時間でいいのかもしれないですね。（鈴木）
  - 25 時間で 1 名で固定でしていただけると助かります。（山本）

LLMの新しいバージョンのリリース後、その日に本番環境のモデルを更新する会社も時々見る。すごいと思っている。

モデルの出力の定性的な評価をする時間はどう考えてもする時間がないはず・・・

新しいモデルだったら基本精度は良いという思い込みのもとやるしかないのか

評価用のデータを用意することも難しい

公開されている音声データと
データの分布が異なるリアルな
会議音声を収集することが必要がある

会議音声の文字起こしと
議事録の形式・内容の
正解データを作成する難しさ

だれがいつ発言をしていたかの
話者のラベリングは
膨大なコストがかかる

宣伝

- NishikaではAI議事録の開発を推進するアプリエンジニアを中心に積極採用中です。詳細はこちら
- Nishika Tech Blogでは毎週のAIの論文情報発信を中心にホットな技術ネタを配信していますので、ぜひチェックしてください！

ご清聴ありがとうございました！