

国語記述式答案自動採点の 現状と実応用に向けて

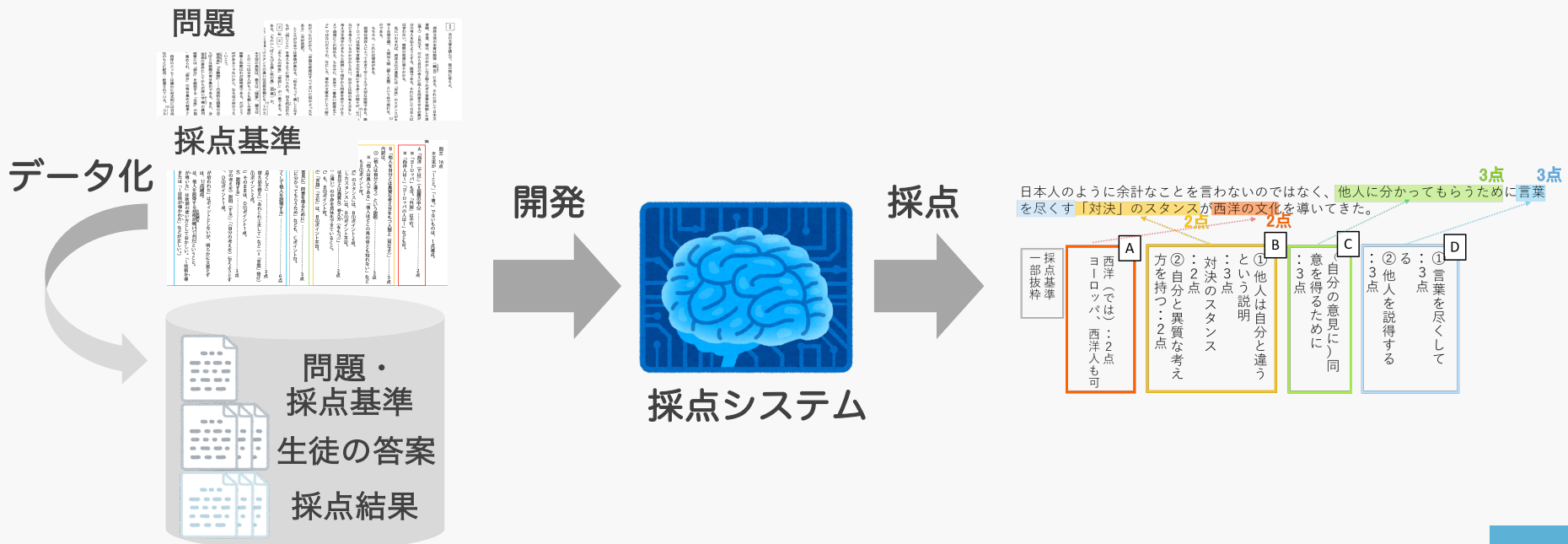
舟山弘晃

東北大学 情報科学研究科 修士課程 /理研AIP 自然言語理解チーム

教育アセスメント×言語処理シンポジウム:
自動採点、英文添削、論述評価の可能性

記述式答案自動採点

- 研究自体は古くから存在も、盛んな分野ではない
 - 原因：利用可能な研究資源(データ)の不足
- 我々の取り組み
 - 教育事業者と提携・生の答案からデータを作成・公開
 - 一部の問題で人間と同等精度の採点を実現 (Mizumoto+, BEA2019)
 - 実応用に向けて、様々な観点から研究(特に解釈性の向上)
 - 基準ごとの採点、採点根拠箇所推定、採点基準の活用、確信度



自然言語処理における自動採点

Essay ScoringとShort Answer scoring(SAS)

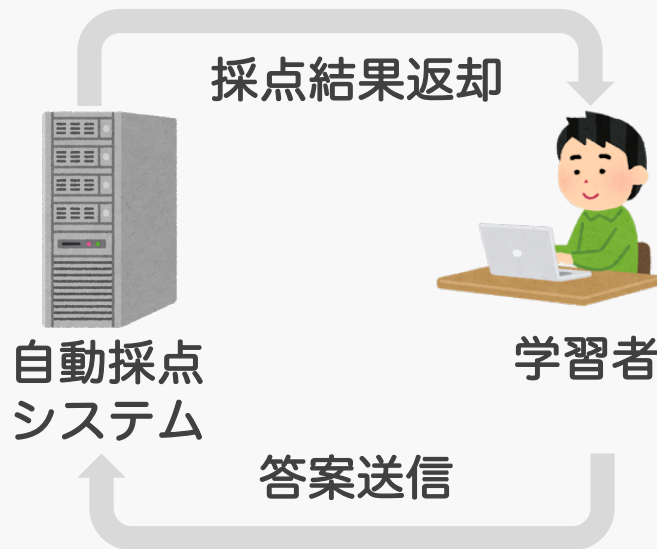
- Essay scoring
 - 国内における小論文に近い
 - 比較的**自由度が高いトピック**に対して、**数百語程度**で論述
 - 文法的正確性・文章の構成・内容の良さなどから**総合的に採点**
- Short Answer Scoring (SAS)
 - 国内における国語・社会・理科における記述式問題に近い
 - 限定的なトピックに対する**理解や知識**を**数十語程度**で記述
 - **記述内容の正当性(採点基準)**に基づいて採点
 - 採点者間で採点結果が一致しやすい
 - **問題ごとに開発(学習)**が必要



本研究のメインターゲットは**Short Answer Scoring (SAS)**

自動採点の利点

- 直ちに結果を返却可能 (即時性)
 - 学習者がその場で答案を修正しながら採点を受けることが可能
- 場所を選ばず利用可能
 - 答案を送信できる電子端末があれば利用可能
 - オンライン学習におけるアウトプットの機会の提供



インタラクティブな学習
の実現

自動採点技術の活用

自動採点を用いた問題集

- 自動採点の特性を利用した問題集
 - 一度開発すれば (翌年度など) 使いまわすことが可能
 - 問題ごとに開発する必要があるものの...
→個別の問題に適した採点AIの開発
- 教育現場における自動採点の実証的な評価
 - 専門家と協議し、より洗練・実用的なAIの開発

自動採点付き問題集



記述式答案自動採点デモ

日本語記述式問題自動採点デモ

次の文章を読んで、後の問いに答えよ。

西洋人は基本的に他人は自分とは異なる考え方をする人間と見なすので、言葉を尽くして他人を説得し、他人に自分の考えを同意させる必要があること。

問：下線部「こうした緊張のスタンスこそが饒舌な西洋文化を導いてきた」とあるが、それはどういうことか。句読点とも七十字以内で説明せよ。

著作権の都合上モザイク処理を行っています

解答入力欄:

西洋人は基本的に他人は自分とは異なる考え方をする人間と見なすので、言葉を尽くして他人を説得し、他人に自分の考えを同意させる必要があること。

0文字

Submit

合計点:

16 / 16

西洋人は基本的に他人は自分とは異なる考え方をする人間と見なすので、言葉を尽くして他人を説得し、他人に自分の考えを同意させる必要があること。

採点基準A

基準

根拠事例

2 / 2

「西洋（では）」（＝話題の中心）・・・2点

※「ヨーロッパ」も可

※「外国」は不可

採点基準B

基準

根拠事例

5 / 5

「他人を自分とは異なる考え方をもち人間と（みなす）」・・・5点

① 他人は自分とは違うという説明・・・3点

※「対決のスタンス」はB②ポイント2点

②（異質な）考え方（を持つ）・・・2点

※B②の違いの中身の具体化

採点基準C

基準

根拠事例

3 / 3

自分の意見に同意を得るために・・・3点

※「他人にわかってもらうため」なども可

採点基準D

基準

根拠事例

6 / 6

「言葉を尽くして他人を説得する」・・・6点

① 「言葉を尽くして」・・・3点

※「手を替え品を替え」や「あれこれ工夫して」などは2点

※「饒舌」そのままは1点

② 「（他人を）説得する」・・・3点

※「（自分の考えを）説明する」や「自分の考えを伝えようとする」は1点

誤字・脱字減点

基準

0 / 1

※誤字・脱字・転記ミス・主述のねじれなどは1点減点

文末減点

基準

0 / 1

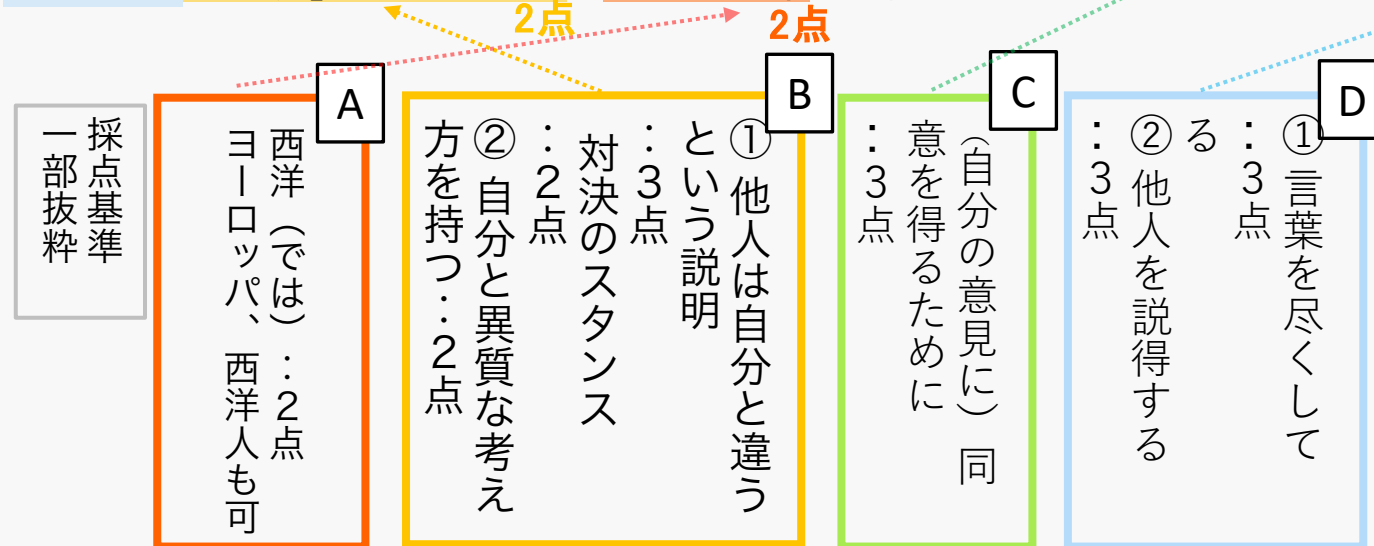
※文末が「こと」、「事」でないものは1点減点

項目自動採点 問題設定

- 問題設定

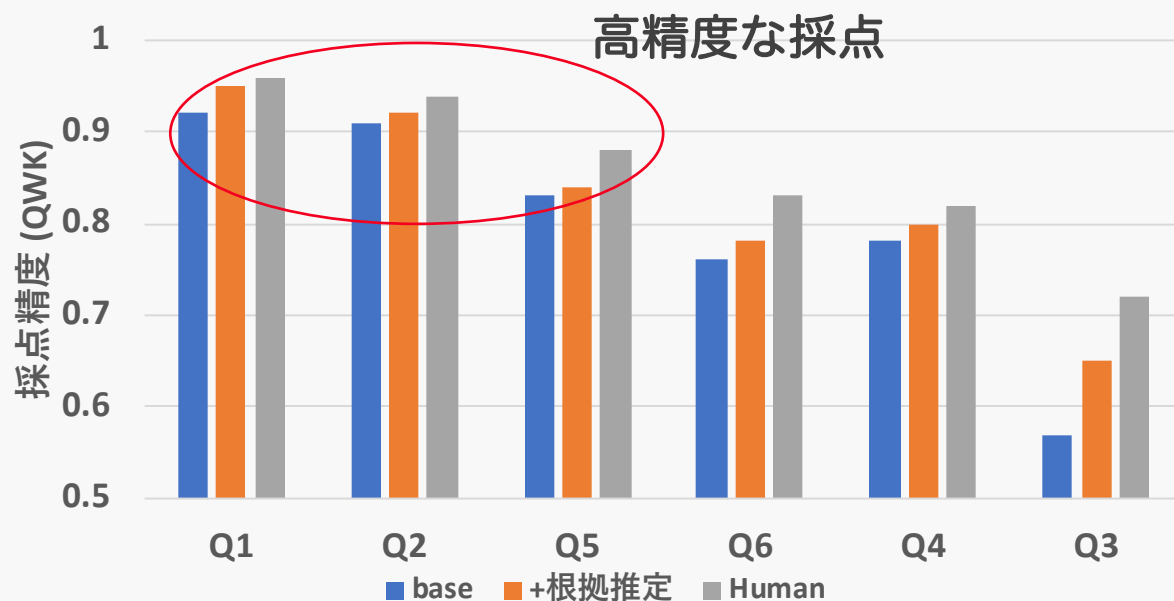
- 解答が特定の情報を含んでいるか判定する情報包含問題を扱う
 - 表現には幅があるためDeep Learningによって柔軟なマッチングを実現
 - 具体的な採点基準が必要

日本人のように余計なことを言わないのではなく、**他人に分かってもらうために言葉を尽くす**「対決」のスタンスが西洋の文化を導いてきた。



項目自動採点 採点精度

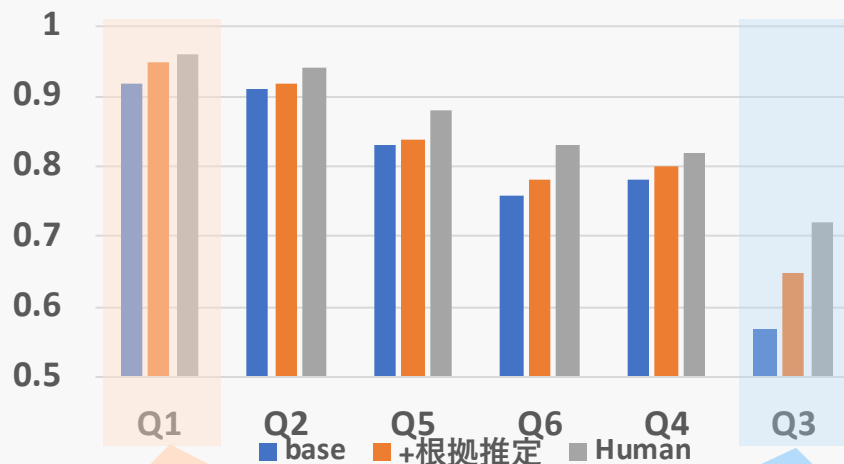
- 採点者間で合意が取れる問題は**高精度で採点が可能**
 - 根拠推定による**精度向上**が可能
 - 学習に必要な解答は**200件程度**



自動採点の活用のためには
人間が採点した時にぶれないような問題であることが重要

項目自動採点 分析

- 自動採点に適した問題とは?



Q1 採点基準より一部抜粋

- ① 「他人は自分と違う」という説明 - 3点
- ② <他人は自分と異質な考え方>を持つ - 2点
 - ①の<違い>の中身が具体化できていること

解答が含むべき内容が**具体的**
＝情報包含問題

Q3 採点基準より一部抜粋

- ① 未練が断ち切れていない事を表現できている - 4点
- ② 困惑した心情の表現 - 2点
 - ネガティブなニュアンスを含む表現を幅広く認める

得点になる**表現が極めて多様**で
自動採点は困難

項目自動採点 データ

- 学習に必要なデータは以下の3点
 - 記述式答案
 - 各答案の採点項目 (基準) ごとの項目点
 - 項目点の根拠箇所
 - 得点データのみでも根拠推定は可能だが精度は下がる

データ例 項目得点および根拠箇所

1	日本人にはない、西洋人ならでは「対決」のスタンスがあり、神や人間や自然との「対決」で生まれた他人を同意させる考え。SPL EOS	A [2]	B [2]	B	C [3]	EOS_ERR [1]
2	西洋人は基本的には他人は自分とは異なる人間と見なす。だから自分の考えに他人を同意させる必要があると考え、	A [2]	B [3]	C [3]		
3	言葉を尽くして考えを伝えようとする。SPL EOS	D [4]	EOS_ERR [1]			
4	こと SPL EOS					
4	西洋の文化の基底には他人に分かってもらうために言葉を尽くして考えを伝えようという「対決」のスタンスがあるということ。SPL EOS	A [2]	C [3]	D [4]	B [2]	

データセット

NIIデータレポジトリ

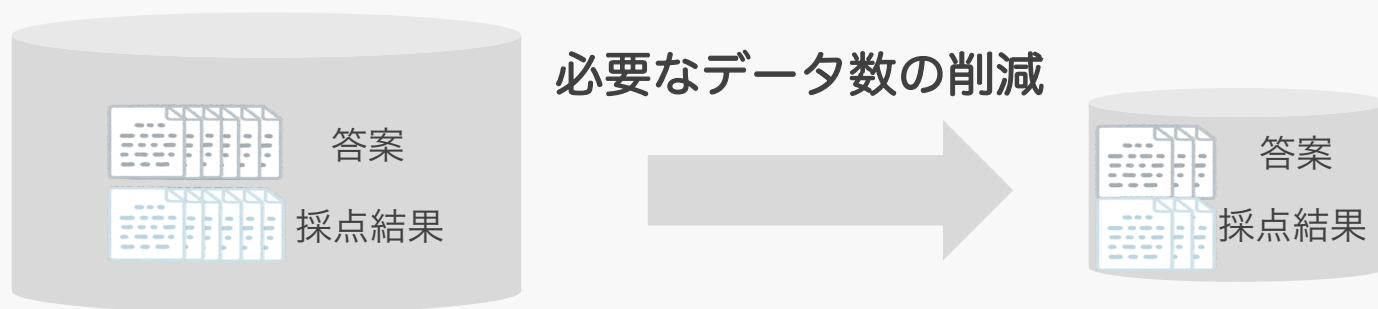


- 国語読解データセット をNIIで公開
 - さらに十数問の拡張を予定
 - 英語を対象としたデータセットも準備中
 - 研究目的で利用可能な初の日本語の大規模データセット

	ASAP-SAS	X-CSD (Mohler+11)	SciEntsBank (Dzikovska+13)	代ゼミ (日本語)	準備中 (英語)
ドメイン	自然科学、 読解など	計算機科学	自然科学	読解	準備中
問題数	10	80	135	9	準備中
問題ごとの 答案数	～2000	28	36	500～2000	準備中
点数	0-2 or 0-3	0-5	2 or 5 段階	0-16	準備中
項目点	X	X	X	✓	✓
採点基準	✓	X	X	✓	✓
根拠箇所	X	X	X	✓	✓

自動採点技術の活用に向けて

- 学習データの確保が最大の課題
 - AIの性能は学習データ量に強く依存
 - 問題ごとに学習データが必要
 - 他研究分野と比べて高いデータ化のコスト
 - 答案収集・書き起こし・採点・アノテーション (根拠付与)
- 日常教育に自動採点を取り入れる場合...
 - 30 (1クラス) ~ 100 (1学年) 件程度で動くAIの開発が望ましい



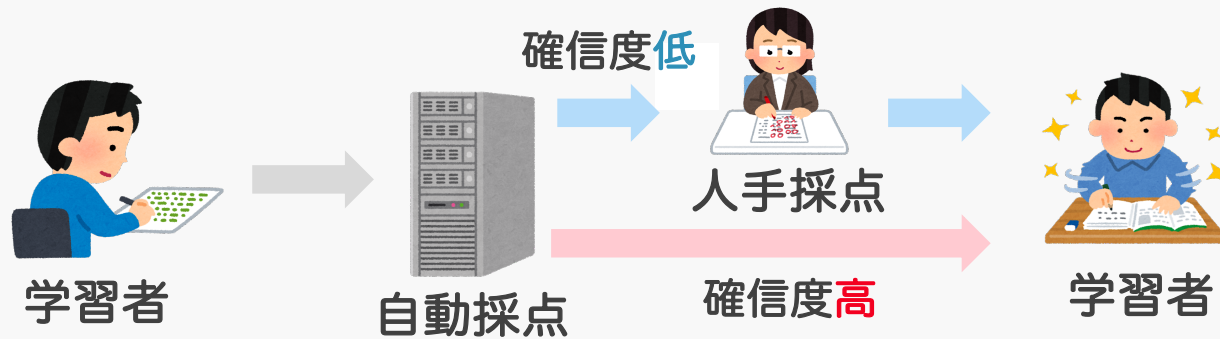
より少量のデータで採点可能な技術の開発

我々のアプローチ

少量データで利用可能な自動採点に向けて

1. 確信度の活用

- 確信度 (信頼度) の開示による解釈性の向上
- 確信度が低い場合は専門家に照会



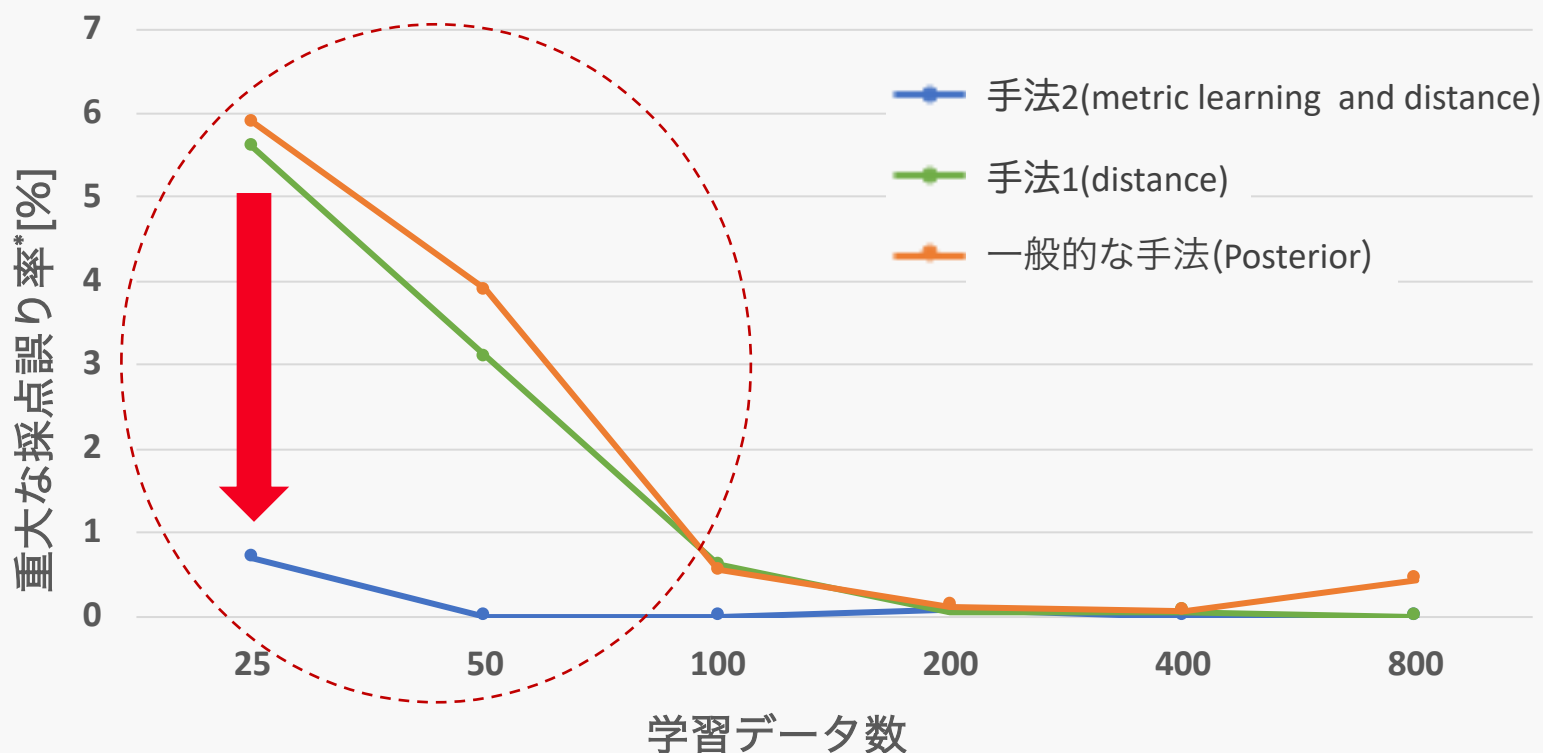
確信度付き自動採点:想定フロー

2. 採点基準 (ループリック) の活用

- 採点基準をAIに組み込むことで必要データ量を削減
- 点数データのみ利用時の根拠推定の精度向上

確信度を用いた自動採点

- 確信度上位80%のテストデータに対する採点誤り

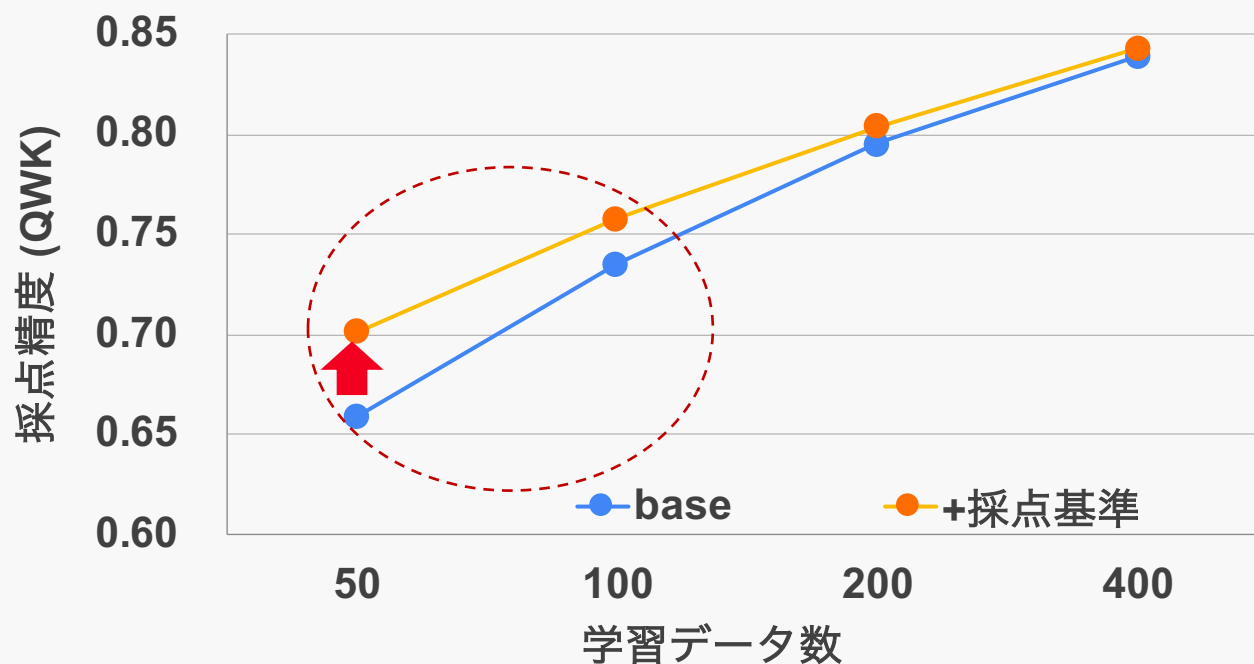


* 配点に対して2割以上の点数幅で採点を誤ることを重大な採点誤りと定義

確信度の活用により
訓練事例数が少ない場合でも重大な採点誤りを除外可能

採点基準 (ルーズブリック) の組み込み

- 採点基準を組み込んだ際の採点精度



採点基準の組み込みによる
訓練事例数が少ない場合の採点精度の向上

おわりに

- 実応用に向けて様々な観点からの研究を実施
 - これらの技術を統合することで、より少ないデータで利用可能な技術の開発を目指す
- さらなる発展に向けて
 - 日本における大規模な自動採点の研究は始まったばかり
 - まだまだ発展途上で、これから更なる発展が予想される
 - 実証実験を行い、実証的知見の積み重ねが重要
 - 学習者・教育の専門家の方からの意見が重要
 - 現在は国語を対象としているが、他の科目にも適用可能
 - 英語に関してはすでに着手（松林から紹介）

ご協力いただける教育事業者・教育関係者の方がいましたらぜひ我々にご連絡ください！