対照研究のための説明可能なAI

［研究代表者］
村脇有吾
京都大学大学院情報学研究科知能情報学専攻

言葉の研究の目的は、人がどのようにして言葉を操っているかを明らかにすることです。人の能力をAIで再現することは、そのための一つの手段です。

京都に行くには何が良いでしょう？

JRか阪急がおすすめです。

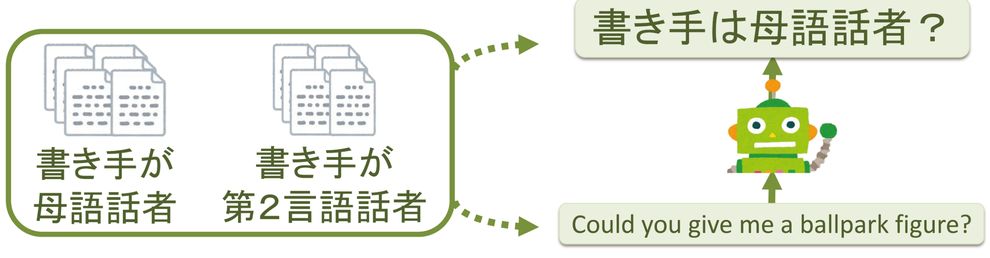
東京に行くには何が良いでしょう？

新幹線を使いましょう。

AIの説明技術は人の説明に転用できるのではないでしょうか？―これが私の基本的なアイデアです。

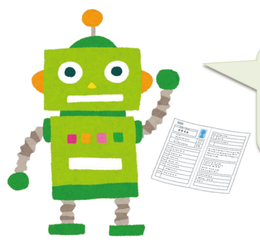
最近のAIは複雑で、何をしているのか開発者にもよくわかりません。それでも、人の脳が言葉を処理する過程が観測できないのと違って、AIなら観測できます。もしAIの処理過程をうまく解析できれば、人の脳の処理過程についても何かわかるのではないでしょうか？

まずは２種類のテキストを見分けるようにAIを訓練します。



次にAIの動作を数学的に解析すると、AIが見分けるための着目した表現が特定できます。

ところで、AIが社会に普及するなかで、AIの意思決定を人が納得できるように説明するための技術が求められています。



残念ながら今回は採用を見合わせていただくこととなりました。

差別じゃないのか！
理由を説明しろ！

例えば、人が英語の母語話者と第２言語話者に長年接していると、「この人は母語話者らしい」といった直感を得ますが、どうしてそう感じたかを人は必ずしも説明できません。そこでAIの出番です！

グループ間の違いを明らかにする研究は対照研究とよばれ、大きな広がりを持っています。例えば、日本語の英訳を元から英語で書かれたテキストと対照すると、英訳しにくい日本語の表現が特定できるかもしれません。

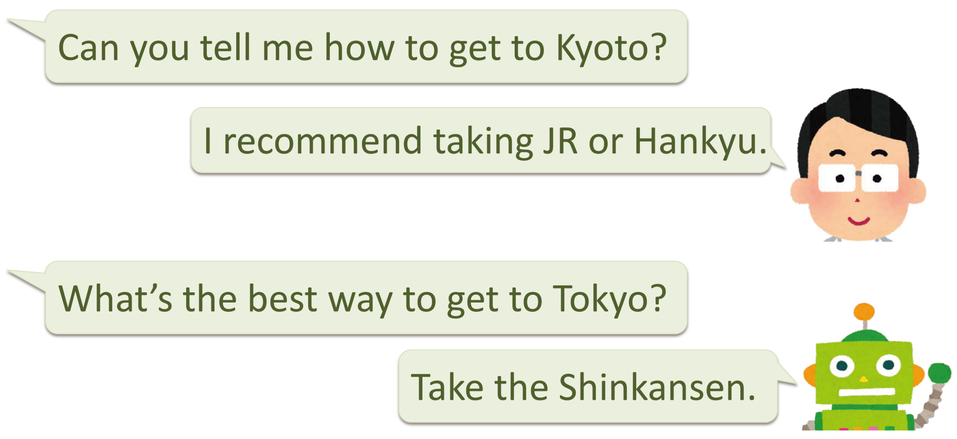
Explainable AI for Contrastive Studies

[Principal Investigator]

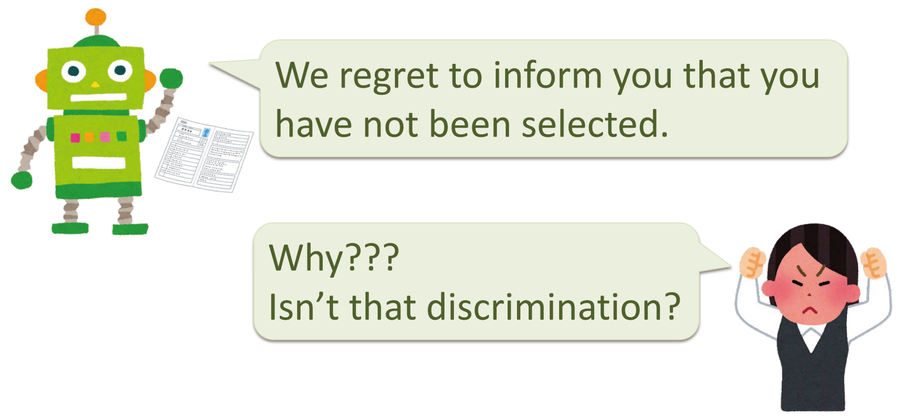
MURAWAKI Yugo

Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

The goal of language science is to unravel the mechanism of human language processing. One promising approach to this challenge is to replicate human ability in AI.



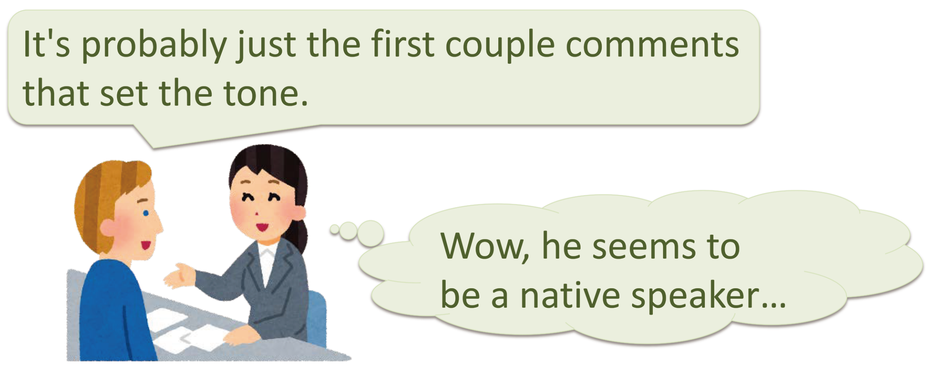
Meanwhile, AI penetration into society is triggering a growing demand for convincing explanations about how AI makes decisions.



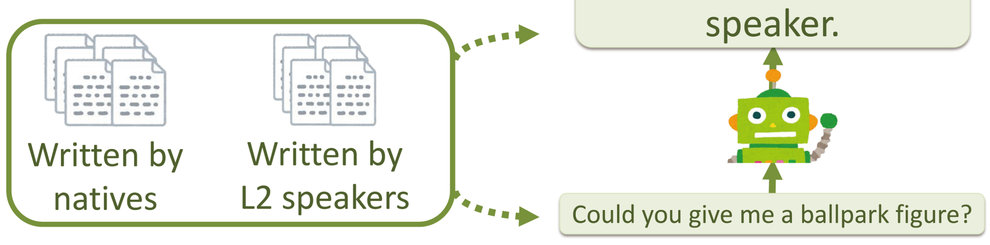
My basic idea is to exploit AI explanation techniques to explain human language faculty.

State-of-the-art AI is so complex that even its developers have trouble understanding what it is doing. Still, unlike human brains, AI allows us to observe how it processes language. If we can successfully analyze AI's language processing, we may be able to learn something about how we humans process language.

For example, years of exposure to native and second language speakers of English gives us a sense of whether someone is a native speaker. However, we cannot always explain why we feel that way. That is where AI comes in!



First, we train the AI to distinguish between the two groups of text.



Next, we mathematically analyze how the AI processes and identifies expressions used by AI to make determinations.

Such contrastive studies have wider potential applications. By contrasting English translations with texts originally written in English, for instance, we may be able to identify Japanese expressions that are difficult to translate into English.