※選択できる教育プログラムについては、入学後、担当教員に確認してください。

データ駆動知識処理研究室
（国立研究開発法人情報通信研究機構）

客員教授：鳥澤　健太郎 torisawa@nict.go.jp
客員准教授：飯田　龍 ryu.iida@nict.go.jp（写真なし）

ビッグデータにどっぷり浸かって、知的で大規模な自然言語処理システムを作り、社会問題の解決に貢献しよう。

研究を始めるのに必要な知識・能力

数学やプログラミングの知識があれば有用ですが、知識がなくても一から勉強できます。むしろ、物事を大胆かつ精密に、またしつこく集中して考えること、そして新しい知識の吸収に貪欲になることが重要です。

研究室の指導方針

本研究室の教員は、情報通信研究機構(NICT)・データ駆動知能システム研究センター (DIRECT)に勤務しており、そこにはACL等の有力国際会議で論文を多数発表している研究者や、プログラマ、言語学者が在籍し、多くの自治体、企業等と連携しつつ、ビッグデータを使った自然言語処理の研究開発を行っています。ビッグデータ、自然言語処理や深層学習等の専門知識や社会貢献に関して、教員だけでなく、これらの人たちとも議論をし、多様な視点を学んでもらえればと思います。また、希望者は自治体の防災訓練等、実社会での応用の現場経験もできます。

この研究で身につく能力

近年、自然言語処理、人工知能の分野では深層学習が注目を集めており、精度の高いニューラルネットワークを設計できる能力が重要になっています。本研究室ではそうした能力を養ってもらう他、ビッグデータを取り扱う大規模自然言語処理システムをデザイン、開発するための大局的視点も学んでもらいたいと思います。これは解決すべき社会問題や自然言語処理、人工知能の将来像を深く考え、それをもとにシステムのアーキテクチャを決定するということです。我々は50名以上の専属の作業者にオリジナルの機械学習用学習データを作ってもらう体制を持っているため、解くべき問題の設計とそれに基づいた学習データの作成から研究を開始することが可能ですが、これは日本ではまれな経験だと思います。

修了生の活躍の場

平成31年度より発足した研究室ですので奈良先端大での修了生はまだいませんが、教員の前職である複数の大学で指導した学生や、過去に情報通信研究機構に在籍し、教員の指導を受けた研究者は、大学教授、大手ネット企業からベンチャーに至るまで様々な職場で活躍しています。

研究内容

A. ビッグデータを用いた知的な対話システムに関する研究

NICT・DIRECTでは、社会に貢献できる自然言語処理システムを目指し、Web40億ページに書かれた知識を用いて雑談も含めた対話を行い、多様な知識を提供できる音声対話システムWEKDA (https://www.nict.go.jp/press/2017/10/24-1.html)や、SNS上の災害情報を分析する対災害SNS情報分析システムDISAANA (https://disaana.jp/)、災害状況要約システムD-SUMM (https://disaana.jp/d-summ/)を開発してきており、DISAANA等については民間企業による商用化が始まっています。(https://jpn.nec.com/press/202006/20200626_01.html) また、近年、大量のテキストで事前学習したいわゆる大規模言語モデルを使うことによって様々な自然言語処理の精度が向上していますが、DIRECTにおいても、代表的な大規模言語モデルであるBERT (パラメータ数は50億個以上に拡大しています。いわゆるBERT-Largeの約15倍のパラメータ数です。) やその派生モデルを、日本語Wikipediaよりも100倍以上大きい、300GB超のWebテキストや、数百枚のGPUさらには、後述する自前開発のミドルウェアRaNNC等も使って構築しており、今後さらに巨大な言語モデルを構築する予定です。現在はそれらの技術をベースに、健康で充実した生活をおくってもらうために高齢者と対話する対話システムMICSUS (https://youtu.be/gCUrC3f9-Go)や、災害時に被災者からの被害情報の収集・提供をスマホ上で行える防災チャットボットSOCDA (https://youtu.be/vvxoMFgd5c8)を研究開発するプロジェクトが進行中ですが、本研究テーマではそれらのプロジェクトには拘らず、ビッグデータを使った知的な対話やディベートを行う技術一般を研究します。研究課題の例としては、「教育目的の対話システムの対話制御技術」、「雑談破綻時の話題・対話戦略修正技術」等が考えられます。

B. ビッグデータを用いた質問応答、仮説生成手法に関する研究

上述の大規模BERTモデルやその派生モデル等を利用し、大規模なWebページ等から知識を取り出す質問応答技術や、取り出された知識からイノベーティブな仮説を生成する技術を研究します。DIRECTの質問応答システムWISDOM X (https://wisdom-nict.jp/#top) は「なぜ」や「どうなる」等の質問にWeb40億ページを元に回答する他、仮説の生成も可能です。科学論文を先取りする仮説を作ることにも成功していますが、ここでの仮説は科学的仮説には限りません。小説のストーリーも一種の仮説と考えられます。対話システムが仮説として生成されたストーリーを勝手に語り始めたら面白いと思いませんか? 研究課題の例としては「長文の回答を要する質問に高精度で回答できる質問応答技術」、「質問応答技術を用いた仮説生成によるストーリー生成技術」等が考えられます。

C. ビッグデータに適用するための自然言語処理基盤技術の研究

これまでに説明してきた技術には自然言語処理の基盤技術である構文解析や意味解析、文脈解析等が必要です。近年ではこれらの基盤技術もこれまでもたびたび言及しているBERT等の巨大言語モデルで精度が大きく向上していますが、DIRECTではそうした言語モデルの中でも一枚のGPUのメモリに格納できない巨大なものを自動的に分割し、複数のGPUで並列に学習／推論させることのできるミドルウェアRaNNCを開発し、上述したように数百GBのテキストを使って、実際に超巨大言語モデルを構築しています。こうした超巨大言語モデルも使った各種の基盤的技術の開発ももちろんwelcomeです。

研究設備

500台以上のサーバー、500枚以上のGPGPU、300億件のWebページ、Twitterに関しては日本語ツイートの10%をリアルタイムで取得。

研究業績・共同研究・社会活動・外部資金など

研究業績については、https://direct.nict.go.jp/publications/、https://direct.nict.go.jpをご覧ください。
外部資金、報道、受賞等については、https://direct.nict.go.jpをご覧ください。なお、NICT DIRECTの成果に関しては、過去5年間で約三百件の報道がなされております。