S-3 学習データの言語学: 機械学習における言葉の選択

Linguistics of training data: language choices in ML model development

佐野大樹（グーグル）

要旨

COVID-19感染予測から、きゅうりの自動選別まで、アルゴリズムの飛躍的な発展・性能の向上により機械学習はいまや我々の生活の様々な側面に浸透している。自然言語処理分野でも、スマートスピーカの開発など、機械学習は言語理解 (natural language understanding) や言語生成 (natural language generation)などに欠かせない「ツール」となっている。

この「ツール」を構築する上で、重要なもののひとつが学習データである。学習データのqualityがツールの性能を左右するようなケースも多く、qualityの高い学習データの構築が、性能の高いツールを開発する上での鍵となる。また、バイアスのある学習データを用いて構築されたツールが、学習データに内在するバイアスを継承してしまい、overfittingや誤ったgeneralization、場合によっては差別などの社会問題を誇張してしまうようなケースもある (e.g., gender bias in machine translation).

本発表では、機械学習を用いたシステムの構築、特に、学習データの構築における言語学的知見や分析方法の応用、および、その役割について説明する。学習データ構築の過程を、(1) デザイン、(2) 作成（生成、拡張もしくは収集）、(3) 分析、(4) 評価に分け、それぞれの過程において言語学がどのように利用できるか説明する。

＜デザイン＞ どのような言語データを学習させる必要があるか
＜作成＞ どのように言語データを収集、もしくは、生成・拡張するか
＜分析＞ どのような言語データが含まれるか、含まれていないか
＜評価＞ 作成した学習データは、タスクにとって有用か

これらの議論を踏まえ、人工知能社会における『学習データの言語学』の重要性について議論したい。

参考文献

Google Cloud (2021) COVID-19 感染予測 (日本版)
https://datastudio.google.com/c/reporting/8224d512-a76e-4d38-91c1-935ba119eb8f/page/ncZpB?s=nXbF2P6La2M [accessed April 2021]

Halliday and Matthiessen (2004) An Introduction to Functional Grammar, Arnold, London.

Melvin Johnson, Senior Software Engineer, Google Research (2020) A Scalable Approach to Reducing Gender Bias in Google Translate, Google AI Blog
https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html [accessed April 2021]

Nithya Sambasivan and Shivani Kapania and Hannah Highfill and Diana Akrong and Praveen Kumar Paritosh and Lora Mois Aroyo (2021) "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI, *SIGCHI*, ACM. https://research.google/pubs/pub49953.pdf [accessed April 2021].

佐藤一憲 (2016) How a Japanese cucumber farmer is using deep learning and TensorFlow, Google Cloud, https://cloud.google.com/blog/products/ai-machine-learning/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow [accessed April 2021].