Generating datasets using people and information on the WWW

01 WWW上のみんな、オラに力を分けてくれ!

WWW上のリソースを活用した機械学習用データ作成手法

どんな研究

WWW上の多種多様、大量のデータに、WWW利用者の力を借りて機械学習用の正解ラベルを付与する仕組みを提供します。既存のWebブラウザにラベリング機能を組み込むことで簡単にラベリングできます。また、サーバと連携して様々な報酬を提供することで、利用者のラベリングを促進します。

どこが凄い

教師あり機械学習で利用する大量かつ高品質な正解ラベル付きデータを作成するには、通常膨大な人的、金銭的リソースが必要です。WWW上のデータ素材に対してWWWの利用者に自発的にラベリングしてもらうというアプローチをとることで、データの準備からラベリングまでを低コストに実現します。

めざす未来

WWW上のデータをWWWの利用者自身が整理していくことができる本技術は、いわばクラウドソーシングの新たな形態を実現します。本技術によって機械学習用の学習データを作成できるようになれば、機械学習を利用した多種多様なサービスをより安価に創出できるようになるでしょう。

AIを支えるのは大量かつ高品質な学習データ

素材の準備

- 良質な素材の選択には時間がかかる。購入は、高額なことも。

素材に対するラベリング

- 作業者への報酬が必要。性能の維持には定期的なラベルの見直しも必要。

WWWを利用している人がWWW上のデータを整理(学習データ化)してくれないかなぁ

World Wide Web (WWW)

世界最大のデータジェネレータ

- さまざまなデータが未整理状態で存在し、また日々更新されている。

第二の生活空間

- 多くの人が日常的にWWW上で活動している(潜在的な作業者)。

WWW横断型ヒューマンコンピュテーション

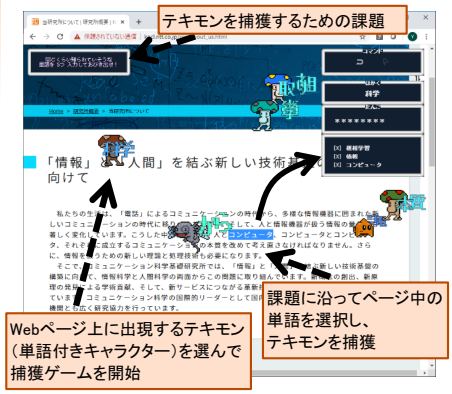
WWW上での活動にちょっとしたプラスアルファを組み込むことで、
WWW上で人が活動していると様々なデータが自動的に整理されていく仕組み

まとめサポーター

複数人で協力して特定のトピックに関する画像を収集していくと、正解ラベルの付与された画像データセットができていく

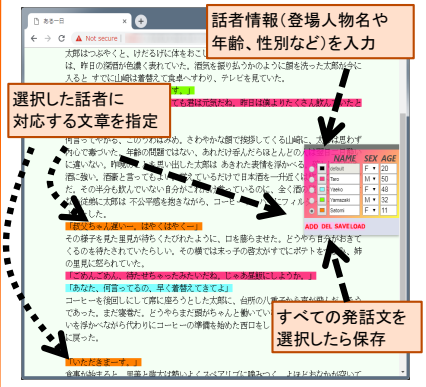
テキストモンスター

Webページに隠れている単語(テキモン)を捕まえ、Webサイトを奪い合うゲームをしていくと、単語に単語親密度が付与されていく



マルチボイスラベラー

小説などを複数の音声合成ボイスで読み上げるためのラベルを付与していくと、談話構造解析用正解データが作成されていく



関連文献

[1] Y. Shirai, Y. Kishino, Y. Yanagisawa, S. Mizutani, T. Suyama, "WWW横断型ヒューマンコンピュテーション," *第27回インタラクティブシステムとソフトウェアに関するワークショップ(WISS2019)*, 2019.

[2] Y. Shirai, Y. Kishino, Y. Yanagisawa, S. Mizutani, T. Suyama, "Building human computation space on the WWW: labeling web contents through web browsers," in *Proc. The seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP2019)*, 2019.

連 絡 先

白井 良成 (Yoshinari Shirai) 協創情報研究部 知能創発環境研究グループ

Email: cs-openhouse-ml@hco.ntt.co.jp

オープンハウス 2020