01 People on the WWW, give us your computation each!

Generating datasets using people and information on the WWW

Abstract

Although contents on the WWW are potentially valuable data as training data for machine learning, they are difficult to use in their current state. Our approach, Browser-based Human Computation (BbHC), offers a cost-effective way to extract desirable data from web contents. BbHC enables people to label various web contents through the web browsers they normally use for web browsing. To accelerate the labeling of data without the inducement of monetary rewards, browser extensions based on BbHC motivate people to continuously engage in labeling tasks through various human computation techniques. We implemented systems based on BbHC to explain how it works. Matome supporter helps us to collect labeled images to create an image classifier. Text monster reduces the cost of annotating word familiarity values for updating a word familiarity database. Multi-voice labeler's purpose is to collect writings with speaker information for natural language processing research.

Deep learning requires much labeled data

Samples to be labeled

- Need budget, if you want to buy them

Labeling task

- Time-consuming, need efforts of many people

Building human computation space on the WWW!

World Wide Web（WWW）

Candidates of samples

- A wide variety of data is stored

Potential workers for labeling tasks

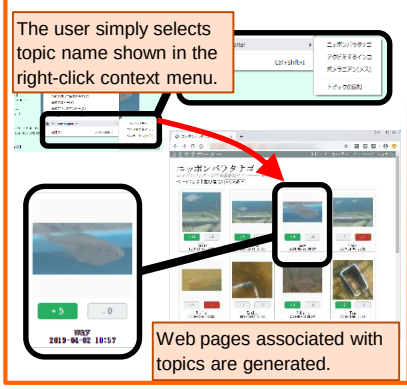
- Many people spends much time on the WWW

Browser-based Human Computation (BbHC)

To collect desirable data from web contents, web browser extensions offer labeling interfaces to users and motivate users to engage in labeling tasks.

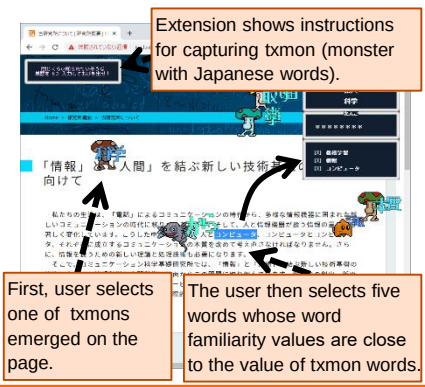
Matome supporter

Users can easily collect images to build web pages that show a collection of images. Collected images are used to update datasets for image classification.



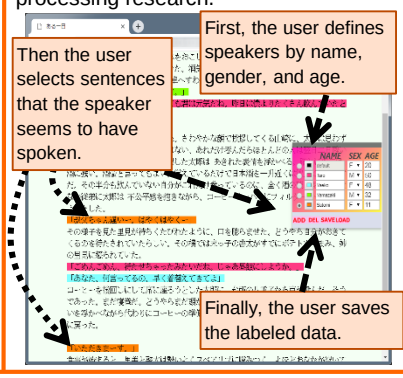
Text monster

Users can enjoy collecting Japanese words which are personified as monsters. The game results are used to update word-familiarity database.



Multi-voice labeler

Users annotate speaker labels to web contents so that smartphones can read them with appropriate voices. Such labels can be used for natural language processing research.



References

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

Contact

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

Learning and Intelligent Systems Research Group, Innovative Communication Laboratory

オープンハウス 2020