Understanding Documents with a Graphical User Interface: Task Design and Model Construction

Akiko Aizawa, National Institute of Informatics

Research Goal

A unified language model that performs diverse document understanding tasks involving language, vision, and actions via a graphical user interface.

Develop tasks that require document comprehension on web-browsers and construct a transformer-based neural model.

First step toward realizing a virtual desktop assistant that can assist and perform tasks requiring document understanding.

- **View photos and fill out the form to answer given questions**
- **Query information using a search interface**
- **Read the document, scroll, locate questions, and input the answer into the form**

**Proposed**

**Assists in working on the desktop**

- Knowledge Graph FAQ bot
- Web QA document structure understanding
- **Document understanding via browser UI**

**Virtual world (block, game)**

**Demo room (robot instruction)**

**Comprehension of instructions in the 3D world**

**Utilizing knowledge represented by textual information**