Dataset 1: chatterbot/english

chatterbot/english is a large dataset of chatbot sentences annotated by topic, category, and more.

Dataset link: https://www.kaggle.com/datasets/kausr25/chatterbotenglish/data

Full github with official chatterbot documentation: https://github.com/gunthercox/chatterbot-corpus/tree/master/chatterbot_corpus/data/english

Dataset 2: ClariQ

The ClariQ dataset contains annotated chatbot dialogs from interactions between humans and machines. The goal of the ClariQ project is to encourage the ML/NLP community to develop chatbots that can detect ambiguity in text and ask quality clarifying questions.

Dataset link: https://github.com/aliannejadi/ClariQ/tree/master/data

Paper link: https://arxiv.org/abs/2009.11352

Paper citation:

ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ)", M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, M. Burtsev, arXiv, 2009.11352, 2020

Dataset 3: Maluuba/frames

Large, annotated dataset of chatbot dialog between AI and human user aimed at aiding the human to complete a task or decision making process, such as booking flights or hotels.

Dataset link: https://github.com/Maluuba/frames

Paper link: https://arxiv.org/abs/1706.01690

Dataset 4: OpenBookQA

OpenBookQA is a large dataset of annotated general-knowledge science questions. The application goal of this dataset is to build assessment tools.

Dataset link: accessing the dataset requires running a Bash script from this repository: https://github.com/allenai/OpenBookQA

Paper link: https://www.semanticscholar.org/paper/Can-a-Suit-of-Armor-Conduct-Electricity-A-New-for-Mihaylov-Clark/1536e8958697c5364f68b2e2448905dbbeb3a0ca