The Ubuntu Chat Corpus for Multiparticipant Chat Analysis

Multiparticipant Chat Corpora

Navy Chat Research
We are investigating techniques for chat analysis to address problems of information overload experienced by US Navy watchstanders. This research requires a suitable chat corpus for experimentation. All Navy chat is unfortunately classified (we want to make our work shareable) so we had to search for a public, unclassified chat corpus.

Multiparticipant Chat Corpora
There is a dearth of publicly-available, large corpora despite multiparticipant chat’s long history. Current corpora are either too small (such as with some labeled corpora) or have an unknown proprietary or privacy status.

Ubuntu

- Ubuntu’s IRC channels allow for real-time technical support
- Ubuntu began using IRC channels in 2004, still currently in use
- Technical, topic-focused

[13:04] <adac> Does external software (software not installed via package manager), even web interfaces go to /opt by default?
[13:04] <jrib> adac: it goes where you want to put it. Customary locations are /usr/local/ and /opt

- All messages are archived, in public domain

Ubuntu Chat Corpus (UCC)

Contents:
- 11 channels (4 in English, 7 in foreign-languages)
- 40+ million messages in total.

Corpora Preprocessing:
- Re-organized file structure
- Removed some system messages – make corpus consistent from beginning to end
- Compressed (2.9GB reduced to 0.6GB)

Benefits:
- Largest publicly-available multiparticipant chat corpus
- All messages in public domain
- Messages are of a technical nature

Drawbacks:
- Messages (initially) unlabeled
- Not suited for social sciences research

Challenge Problem #1 – Intelligent Word Highlighting

Problem:
- Multiparticipant chat clients offer limited highlighting capabilities
- User can enter set of words (or regular expressions) to be highlighted
- Highlighting will fail in cases of misspelled words, abbreviations, or synonyms

Goal:
- Create automated techniques for finding words related to a user’s interests from past history of chat messages

Our status:
- Annotated a corpus for evaluation, subset of UCC labeled for relation to “Unity”
- Created an unsupervised algorithm to learn word relations from unlabeled chat
- Our algorithm outperformed a baseline approach (similar to state-of-the-art chat clients)

Publication: (Uthus & Aha, FLAIRS-13): Extending Word Highlighting in Multiparticipant Chat

Challenge Problem #2 – Intelligent Bots

Problem:
- There are bots in the IRC channels that have access to databases of factoids, which are often used for answering frequently-asked questions
- Bots must be manually invoked by experts to answer a user’s question

Goal:
- Automate the bots to answer questions they can confidently answer
- Allow experts to focus their attention on difficult, less-common questions

Our status:
- Annotated a corpus for evaluation, 4000+ questions labeled from UCC
- Empirical studies show a bot can answer some questions accurately

Publication: In submission

Challenge Problem #3 – Automatic Chat Summarization

Problem:
- Many years of messages are archived but are not being reused (to our knowledge)
- Difficult to search for past solutions (i.e., to technical problems)

Goal:
- Automatically extract factoids, which are answer summaries to FAQs

Our status:
- Currently investigating techniques for this problem
- We will use human-authored factoids as gold standards for evaluations

Availability
The Ubuntu Chat Corpus (and annotated subsets) are available at: http://daviduthus.org

Navy Center for Applied Research in AI (Code 5510) - Washington, DC
Email: david.uthus.ctr@nrl.navy.mil
WWW: http://daviduthus.org