The Ubuntu Chat Corpus for Multiparticipant Chat Analysis

David C. Uthus[1,2] David W. Aha[2]

[1]National Research Council Postdoctoral Fellow
[2]Naval Research Laboratory

AAAI Spring Symposium on Analyzing Microtext
March 2013

Navy Research and Chat

We needed a chat corpus for Navy-related research.

- Navy chat is classified.
- Want to make research sharable.

Multiparticipant Chat Corpora

Lack of suitable corpora for multiparticipant chat analysis despite long history of chat:

- MUDs began in 1970s.
- IRC created in 1988.

Currently-available (labeled) corpora:

- NPS Chat Corpus (Forsyth & Martell, 2007; Lin, 2007)
- `#LINUX` corpus (Elsner & Charniak, 2010)
- `#IPHONE`/`#PHYSICS`/`#PYTHON` corpus (Adams, 2008)

Also many archives of chat logs, proprietary status unknown.

[1]

[1] The authors are not affiliated with Ubuntu.

This slide has been removed due to the 2013 US Budget Sequestration

♡Obama, Reid, & Boehner

Ubuntu IRC Technical Support Channels

Community-run, Ubuntu-supported real-time support.

```
[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
```

Began with one channel in 2004, grown to multiple support channels (including foreign-language support) since then.

Ubuntu Chat Corpus

Structure:

- 11 channels, ranging from 550K – 26M messages each.
- 4 English channels, 7 language- & country-specific channels.

Preprocessing:

- Re-organized file structure – easier to use for research.
- Some system messages removed for consistency.
- Compressed (2.9GB reduced to 0.6GB).

Publicly available at `http://daviduthus.org`

Ubuntu Chat Corpus

Benefits:

- Largest publicly-available chat corpus
- Topic-focused, technical chat messages
- Messages in public domain

Drawbacks:

- All data unlabeled
- Not suited for social sciences research

Conclusions

The Ubuntu Chat Corpus addresses the lack of a large, publicly available multiparticipant chat corpora.

- Public domain messages
- Largest chat corpus available
- Technical chat, not social

Come see our poster for ideas about research topics for this corpus!

Thank you!

References I

Adams, P. H. (2008). Conversation Thread Extraction and Topic Detection in Text-Based Chat. Master's thesis, Naval Postgraduate School.

Elsner, M., & Charniak, E. (2010). Disentangling chat. *Computational Linguistics*, *36*(3), 389–409.

Forsyth, E. N., & Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *Proceedings of the International Conference on Semantic Computing*, pp. 19–26. IEEE Computer Society.

Lin, J. (2007). Automatic Author Profiling of Online Chat Logs. Master's thesis, Naval Postgraduate School.