The Ubuntu Chat Corpus for Multiparticipant Chat Analysis

David C. Uthus
NRC/NRL Postdoctoral Fellow
Washington, DC 20375
david.uthus.ctr@nrl.navy.mil

David W. Aha
Naval Research Laboratory (Code 5514)
Washington, DC 20375
david.aha@nrl.navy.mil

Abstract

We present the Ubuntu Chat Corpus as a data source for multiparticipant chat analysis. This addresses the problem of the lack of a large, publicly suitable corpora for research in this medium. The advantages of using this corpus for research is its large number of chat messages, its multiple languages, its technical nature, and all of the original chat messages are in the public domain.

Introduction

Multiparticipant chat is a form of chat where multiple participants are conversing synchronously through text communications. Examples of multiparticipant chat include Internet Relay Chat (IRC), virtual game lounges (e.g., Battle.net, Steam), game environments (e.g., MUDs, MMORPGs), and collaborative learning environments. Growing interest in the past decade has looked at problems such as thread disentanglement (Elsner and Charniak 2010), topic detection (Durham 2009; Trausan-Matu et al. 2007), author profiling (Lin 2007; Köse, Özyurt, and İkibaş 2008), and message attribute identification (Dela Rosa and Ellen 2009; Wu et al. 2005). One thing this research area has been missing though is a large, public corpus of chat messages. Having such a corpus would allow for better comparison of different techniques, standardization of evaluations, and make it easier for researchers to enter the field.

We describe the Ubuntu Chat Corpus as a data source of research for multiparticipant chat analysis. This corpus consist of messages from Ubuntu's IRC support channels. In this paper, we first describe how we constructed this corpus, followed by how it compares with other chat data sources. We then conclude by proposing some open research problems that can be investigated using this corpus.

Background

Chat has not had a large corpora available for public use despite it being an old medium – MUDs began in the 1970s and IRC was created in 1988 (Herring in press; Reid 1991). There are some comparatively small, annotated corpora being used for current chat research, such as the NPS Chat Corpus (Forsyth and Martell 2007; Lin 2007), the #LINUX corpus (Elsner and Charniak 2010), and the #IPHONE/#PHYSICS/#PYTHON corpus (Adams 2008). For many other research investigations though, the authors either used archives which have unknown copyright status or used self-collected data which were not made publicly available, making it difficult to comparatively evaluate different techniques.

Corpus Description

Ubuntu, a Linux-based operating system, has multiple IRC channels (found on the freenode IRC network) for technical support and development coordination. Ubuntu started using IRC in 2004 with one channel, #ubuntu (which is still their primary support channel), and has since then expanded to multiple channels for more specific topics as well as support in non-English languages. All messages are logged and kept in a public archive at http://irclogs.ubuntu.com/.

We created a corpus from the logs between 2004-07-05 until 2012-10-17 (the day before the release of Ubuntu 12.10). We selected eleven frequently-used channels from the archive, including seven non-English channels, which are listed in Table 1. We removed all system messages (e.g., users entering or leaving a channel) except for messages which indicate a user changing their nickname. We did this to make the logs consistent – Ubuntu originally recorded all system messages but later recorded only the nickname changes. All files in the corpus are encoded in UTF-8, and the corpus was compressed from 2.9GB to 0.6GB. The corpus is available at http://daviduthus.org/.

Figure 1 shows the volume of messages over time in the channel #ubuntu, visualizing the cyclical pattern of traffic seen in Ubuntu's support channels. As can be seen, the number of messages spikes every six months, which coincides with Ubuntu's bi-annual updated release. The greatest peak was on 2006-05-27, when there were 58 900 messages recorded that day, or 0.7 messages per second.

An unfortunate trend seen in the graph is that the volume of messages has decreased. This downward trend does not diminish the validity of using this corpus. freenode, which hosts IRC channels for many open-source projects, has had an increasing number of users in recent years (freenode 2012). Thus research on this corpus benefits other open-source projects that use IRC for their technical support.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Ubuntu Chat Corpus for Multiparticipant Chat Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Code 5514,4555 Overlook Ave., SW,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
in AAAI Spring Symposium on Analyzing Microtext, Stanford, CA, 25-27 Mar 2013.

14. ABSTRACT
We present the Ubuntu Chat Corpus as a data source for multiparticipant chat analysis. This addresses the problem of the lack of a large, publicly suitable corpora for research in this medium. The advantages of using this corpus for research is its large number of chat messages its multiple languages, its technical nature, and all of the original chat messages are in the public domain.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

Channel	Number of Messages	Number of Users	Avg Msg Length	First Logged	Description
#ubuntu	26 360 715	529 882	57.6	2004-05-07	Ubuntu's primary support
#kubuntu	5 963 258	100 411	47.6	2005-03-31	Kubuntu (Ubuntu with KDE) support
#ubuntu-devel	2 112 074	12 140	53.7	2004-10-01	Developmental team coordination
#ubuntu+1	1 621 680	26 805	52.6	2007-04-04	Developmental versions' support
#ubuntu-cn	1 641 416	11 162	21.7	2010-11-04	Support for Mainland China
#ubuntu-ru	883 662	8 320	40.6	2010-11-04	Support for Russia
#ubuntu-br	649 969	6 725	34.4	2010-11-04	Support for Brazil
#ubuntu-es	646 675	9 020	41.3	2010-11-04	Support for Spanish speakers
#ubuntu-it	645 375	10 316	47.0	2010-11-04	Support for Italy
#ubuntu-pl	635 873	3 467	33.1	2010-11-04	Support for Poland
#ubuntu-se	550 013	2 456	45.2	2010-11-04	Support for Sweden

Table 1: Description of the channels included in the Ubuntu Chat Corpus. The average message length is defined as the average number of characters excluding the timestamp and user's nickname.

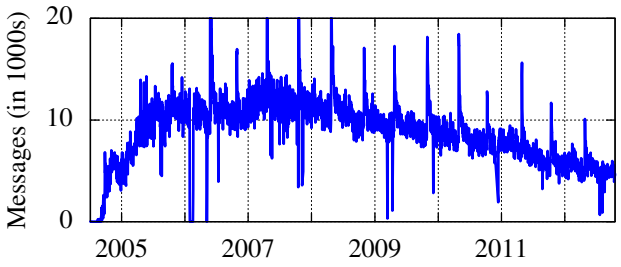

Figure 1: The daily volume of messages in #ubuntu.

Comparison with Other Chat Data

There are four reasons for using this chat corpus compared to other collections of multiparticipant chat logs.[1] First is the corpus size – this is the largest collection of publicly-available chat logs that we are aware of.

Second, the original messages from the archive are in the public domain, bypassing many legal issues. This is both stated on Ubuntu's website[2] and it is sent as a message whenever a user enters one of their logged channels. We have not found any other set of chat logs that have been explicitly declared as public domain.

Third, the logs contain technical discussions, which allows for research that is applicable to other technical domains (e.g., business, online courses, collaborative learning, military command and control). There is one disadvantage to this – these channels have less social chat than would be seen in a non-technical chat channel.

Finally, the corpus contains channels in languages other than English, yet the channels cover the same general topics as discussed in the English channels. While there is no exact match between messages or time-specific topics in the different channels, one can assume popular topics being discussed in the main support channel (which is English-only) will probably also be popular in the non-English channels.

[1]These four claims come from our difficult experience of finding a suitable chat corpus for our research.

[2]https://help.ubuntu.com/community/InternetRelayChat/

Challenge Research Problems

We now describe some challenge research problems that can be investigated using this chat corpus. We focus on problems that have received minimal or no research attention in a multiparticipant chat domain. To overcome these problems, either techniques from other domains will need to be adapted or new techniques will need to be created for this domain.

Intelligent Word Highlighting

Word highlighting helps users find messages of interest. Unfortunately, current state-of-the-art word highlighting for chat clients is rather simple – users enter a list of words they want highlighted, and the client will only highlight these specific words. There are many problems with this: words can be misspelled or abbreviated; words can be falsely highlighted due to lack of context awareness; or relevant words may be missed since it is difficult to predict all the words that one might need highlighted. In addition, different channels (such as in IRC) can have different meanings for the same word, e.g., the term "unity" spoken in an Ubuntu-related channel would refer to Ubuntu's new user interface while it would usually not have a special meaning in a non-Ubuntu-related channel.

Given this difficulty of finding relevant messages in chat, techniques are needed to aid users in filtering the messages. For example, techniques that can suggest related words to those which a user would want highlighted could aid users in finding more relevant messages and easily integrate with current chat clients. So far, a few researchers have investigated this problem from a military perspective, with the goal of reducing information overload for military personnel who use chat for command and control communications (Berube et al. 2007; Budlong, Walter, and Yilmazel 2009; Dela Rosa and Ellen 2009).

Intelligent Bots

Ubuntu uses bots to aid with running their IRC channels. One of these bots, `ubottu`, contains a collection of factoids (short messages), which can be used to answer other peoples' questions. While this is evidently helpful (`ubottu` has generated the most messages in this corpus), the question-answering process is unfortunately done manually; an experienced user must direct the bot as to which question to answer and which factoid to use.

Essentially, the challenge is how to create an intelligent bot that could confidently answer common questions correctly while allowing more expert users to answer questions beyond its capabilities. An even more difficult problem would be to create a bot that can learn answers to new types questions, such as when new software has been introduced in Ubuntu. Some research has investigated creating intelligent agents in a multiparticipant chat domain, such as Cobot (Isbell et al. 2006). These agents, while limited in conversation ability, can provide a starting point for more intelligent, interactive bots in such a domain.

Automatic Chat Summarization

As previously shown, there are many years worth of chat messages and knowledge archived by Ubuntu which, as far as we know, is not being reused outside of human-made factoids for `ubottu`. Summarization techniques leveraged to summarize answers to frequent questions would be beneficial, as this would then allow for this knowledge to be reused. This can also be used by intelligent bots to create answers to new types of questions. An advantage of this chat corpus is that there are already human-authored summaries in the form of factoids that can be used as gold standards for evaluations.

There has been little research on automatic multiparticipant chat summarization. Zhou and Hovy (2005) investigated summarizing chat messages with extractive methods to create summaries similar to human-made summaries. We recently described our goals for investigating how to summarize conversation threads of chat messages (Uthus and Aha 2011).

Multi-Language Techniques

Most research on multiparticipant chat has used chat logs whose messages are in only a single language, with most focusing on English. This corpus provides a great resource for investigating techniques on non-English languages and for investigating techniques which are language-independent, such as for thread disentanglement.

Another research problem on multiple languages in chat is translating chat messages. `#ubuntu` is visited by far more users than any of the non-English channels, so it is easy for a user to receive help if one is fluent in English. For those who are not, there might not be many expert users who write in their native language, which can then make it difficult to receive any help. Machine translation could then help overcome this problem. So far, there has been some limited work focused on multiparticipant chat translation (Calefato, Lanubile, and Minervini 2010; Yamashita et al. 2009; Yoshino and Ikenobu 2010), but these studies only examined users chatting in small group settings on constrained tasks.

In relation to intelligent bots, this corpus can be used to detect non-English messages in Ubuntu's English support channels. This can then aid in directing users to more appropriate channels. Recent similar work has been reported on detecting the language of tweets for creating language-specific Twitter collections (Bergsma et al. 2012), which can be used as a starting point due to some of the shared similarities between microblogs and chat.

Conclusions

We have presented the Ubuntu Chat Corpus as a data source for research on multiparticipant chat analysis. It has many benefits that make it useful for research in this medium: its large size, its public domain status, its technical discussions, and it contains chat logs in non-English languages. We have also described some challenging problems in multiparticipant chat analysis that have received little research attention and which would be suitable to investigate with this corpus.

Acknowledgments

Thanks to NRL for funding this research. David Uthus performed this work while an NRC postdoctoral fellow located at the Naval Research Laboratory. The views and opinions contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of NRL or the DoD.

References

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

Bergsma, S.; McNamee, P.; Bagdouri, M.; Fink, C.; and Wilson, T. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, 65–74. Association for Computational Linguistics.

Berube, C. D.; Hitzeman, J. M.; Holland, R. J.; Anapol, R. L.; and Moore, S. R. 2007. Supporting chat exploitation in DoD enterprises. In *Proceedings of the International Command and Control Research and Technology Symposium*. CCRP.

Budlong, E. R.; Walter, S. M.; and Yilmazel, O. 2009. Recognizing connotative meaning in military chat communications. In *Proceedings of Evolutionary and Bio-Inspired Computation: Theory and Applications III*. SPIE.

Calefato, F.; Lanubile, F.; and Minervini, P. 2010. Can real-time machine translation overcome language barriers in distributed requirements engineering? 257–264. IEEE Computer Society.

Dela Rosa, K., and Ellen, J. 2009. Text classification methodologies applied to micro-text in military chat. In *Proceedings of the International Conference on Machine Learning and Applications*, 710–714. IEEE Computer Society.

Durham, J. S. 2009. Topic Detection in Online Chat. Master's thesis, Naval Postgraduate School.

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

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

freenode. 2012. History and growth. http://freenode.net/history.shtml.

Herring, S. C. in press. Relevance in computer-mediated conversation. In Herring, S.; Stein, D.; and Virtanen, T., eds., *Handbook of the Pragmatics of Computer-Mediated Communication*. Mouton de Gruyter.

Isbell, C. L.; Kearns, M.; Singh, S.; Shelton, C. R.; Stone, P.; and Kormann, D. 2006. Cobot in LambdaMOO: An adaptive social statistics agent. *Autonomous Agents and Multi-Agent Systems* 13(3):327–354.

Köse, C.; Özyurt, O.; and İkibaş, C. 2008. A comparison of textual data mining methods for sex identification in chat conversations. In *Proceedings of the Fourth Asia Information Retrieval Conference on Information Retrieval Technology*, 638–643. Springer-Verlag.

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

Reid, E. M. 1991. Electropolis: Communication and community on Internet Relay Chat. University of Melbourne Honours Thesis.

Trausan-Matu, S.; Rebedea, T.; Dragan, A.; and Alexandru, C. 2007. Visualisation of learners' contributions in chat conversations. In *Proceedings of the Workshop on Blended Learning*, 217–226. Pearson.

Uthus, D. C., and Aha, D. W. 2011. Plans toward automated chat summarization. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, 1–7. ACL.

Wu, T.; Khan, F. M.; Fisher, T. A.; Shuler, L. A.; and Pottenger, W. M. 2005. Posting act tagging using transformation-based learning. In Young Lin, T.; Ohsuga, S.; Liau, C.-J.; Hu, X.; and Tsumoto, S., eds., *Foundations of Data Mining and Knowledge Discovery*, volume 6 of *Studies in Computational Intelligence*. Springer Berlin / Heidelberg. 321–331.

Yamashita, N.; Inaba, R.; Kuzuoka, H.; and Ishida, T. 2009. Difficulties in establishing common ground in multiparty groups using machine translation. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, 679–688. ACM.

Yoshino, T., and Ikenobu, K. 2010. Availability of multilingual chat communication in 3D online virtual space. In Ishida, T., ed., *Culture and Computing*, volume 6259 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 28–40.

Zhou, L., and Hovy, E. 2005. Digesting virtual "geek" culture: The summarization of technical Internet Relay Chats. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 298–305. ACL.