Extending Word Highlighting in Multiparticipant Chat

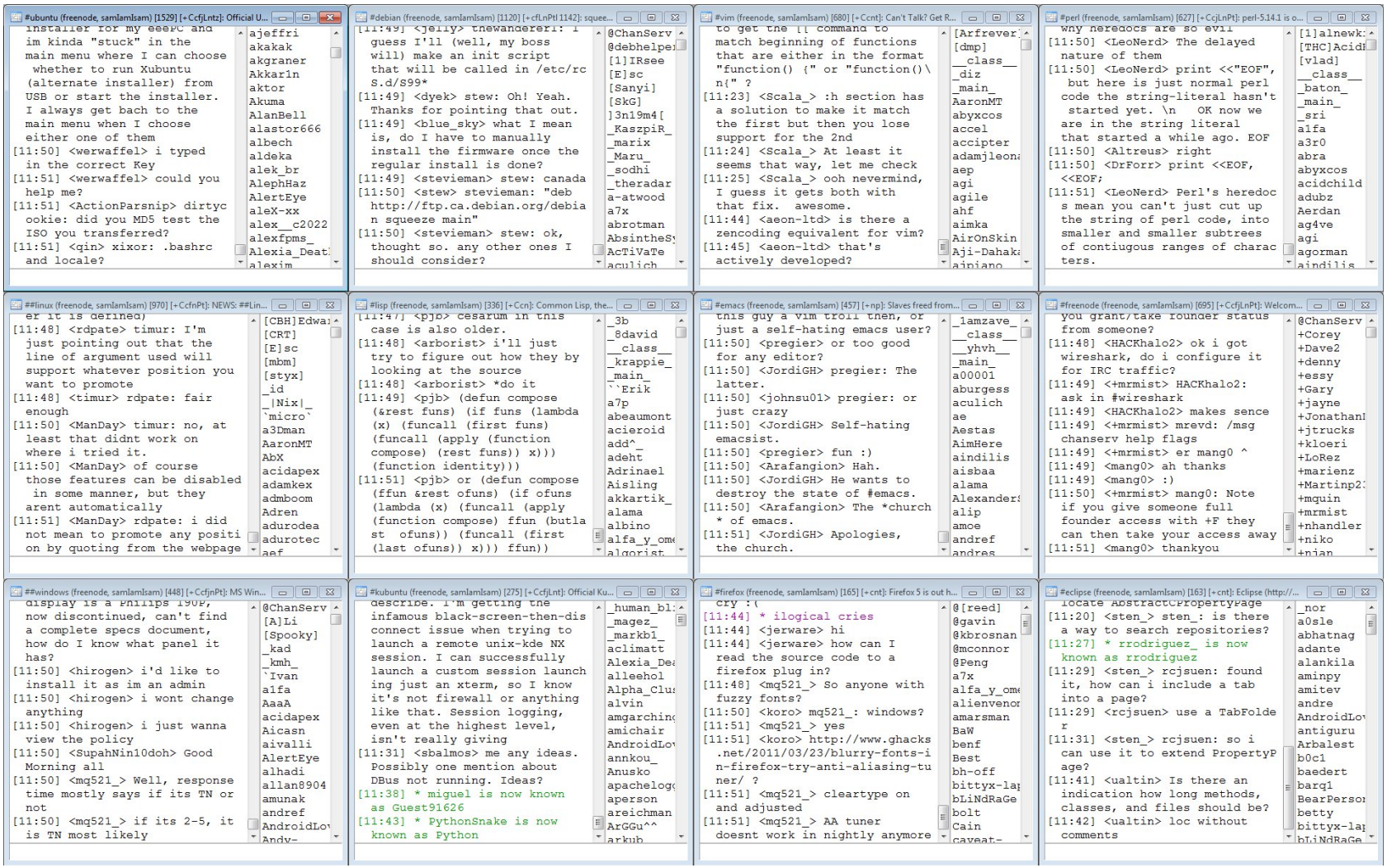
David C. Uthus[1,2] David W. Aha[2]

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

FLAIRS, May 2013

In The Navy. . .

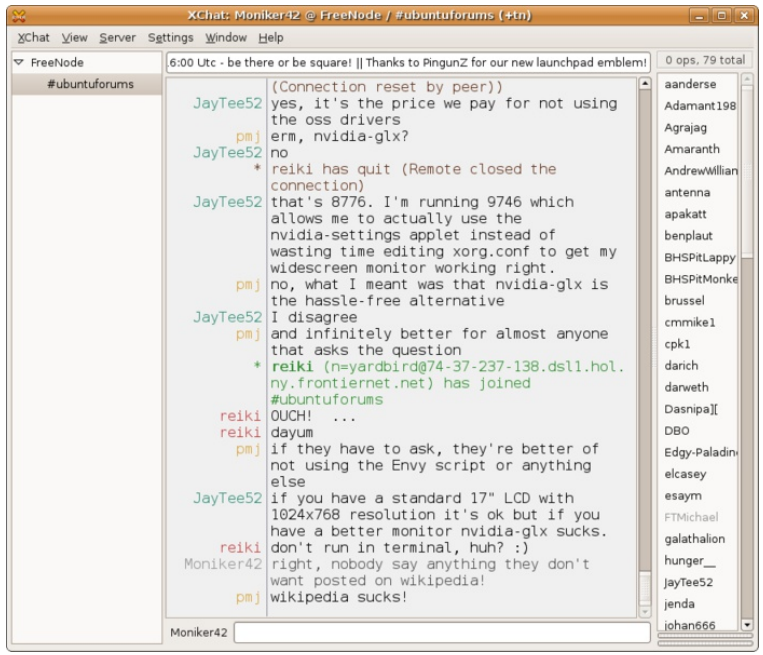
In The Navy…


#ubuntu (freenode, samlamlsam) [1529] [+CcfjLntz]: Official U...
#debian (freenode, samlamlsam) [1120] [+cfLnPtl 1142]: squee...
#vim (freenode, samlamlsam) [680] [+Ccnt]: Can't Talk? Get R...
#perl (freenode, samlamlsam) [627] [+CcjLnPt]: perl-5.14.1 is o...
##linux (freenode, samlamlsam) [970] [+CcfnPt]: NEWS: ##Lin...
#lisp (freenode, samlamlsam) [336] [+Ccn]: Common Lisp, the...
#emacs (freenode, samlamlsam) [457] [+np]: Slaves freed from...
#freenode (freenode, samlamlsam) [695] [+CcfjLnPt]: Welcom...
##windows (freenode, samlamlsam) [448] [+CcfjnPt]: MS Win...
#kubuntu (freenode, samlamlsam) [275] [+CcfjLnt]: Official Ku...
#firefox (freenode, samlamlsam) [165] [+cnt]: Firefox 5 is out h...
#eclipse (freenode, samlamlsam) [163] [+cnt]: Eclipse (http://...

Multiparticipant Chat Client Technology

State-of-the-art (e.g., mIRC, XChat, Pidgin, Konversation) is simplistic:

- User enters words, client highlights only these words
- Some allow regular expressions, but still fragile



Related Work

Three investigations on finding important messages in chat (all military-related):

- Berube, Hitzeman, Holland, Anapol, and Moore (2007)
- Budlong, Walter, and Yilmazel (2009)
- Dela Rosa and Ellen (2009)

Goal

Extend chat client highlighting capability in order to highlight more relevant **WORDS** to improve a users' ability to find relevant messages using unsupervised learning methods.

Problem Description

Given:

- U – Set of unlabeled messages (for training)
- L – Set of labeled messages (for testing)
- W – Set of user-specified words (USWs)

Task:

- Learn model such that, when given a word $w \in W$, the model will return a set of related words R
- For each message $l \in L$, mark positive any message containing a word $r \in R \cup \{w\}$, otherwise mark negative

Objective:

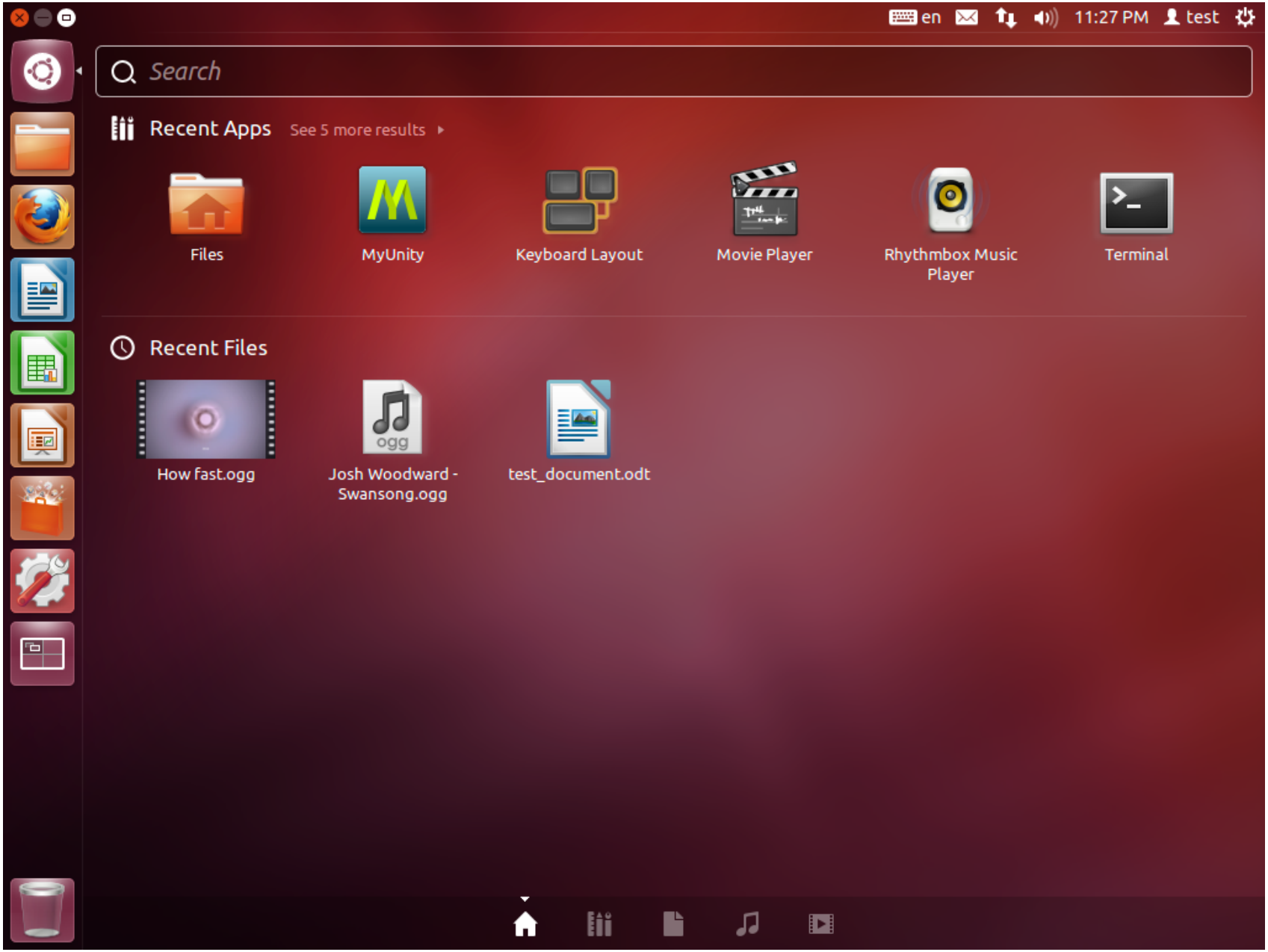
- Find set of words that maximizes coverage of positive labels and minimizes coverage of negative labels in L

Ubuntu IRC Technical Support

Corpus

Used a subset of chat logs from the Ubuntu Chat Corpus (Uthus & Aha, 2013):

- Training set – seven days of unlabeled messages (28 April to 4 May 2012)
- Test set – one day of manually labeled messages (5 May 2012) with binary label of whether they are related to Unity


en
11:27 PM
test
Search
Recent Apps
See 5 more results
Files
MyUnity
Keyboard Layout
Movie Player
Rhythmbox Music Player
Terminal
Recent Files
How fast.ogg
Josh Woodward - Swansong.ogg
test_document.odt

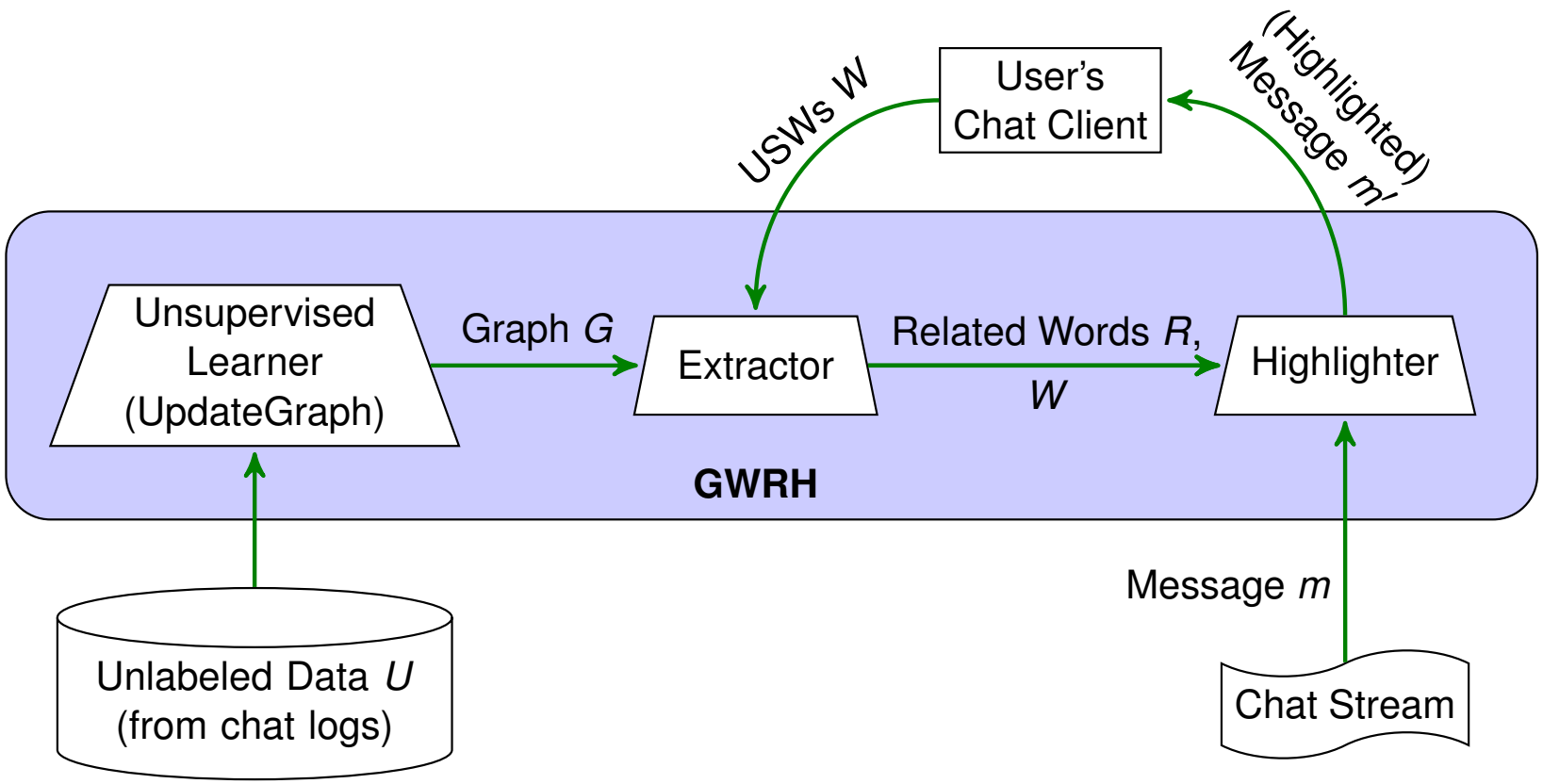
Corpus – Difficulties

```
[05:40] <daddy> is there a way to change 11.04 interface
back to 10.10
[05:40] <DrFrankenstein> daddy:  launching programs from a
drop down menu instead of the screen with the icons?
[05:40] <Soupermanito> yes, log out and choose clasic
interface at the log in menu
```

Corpus – Difficulties

[01:24] <bible-boy> anything else like unity that i can use on 11.04
[01:24] <bible-boy> cause i like unity
[01:24] <TomRone> bible-boy, install fluxbox desktop environment with synaptic and give that a shot or lxde perhaps. you use the login manager to choose which environment to use
[01:25] <lapion> bible-boy, why can't you use unity ?
[01:25] <bible-boy> where can i download that because i have it installed on a pc without internet. Right now im running Ubutnu off a live cd
[01:25] <bible-boy> oh i cant use it because i only have 512mb memory

GWRH – General Framework



Baselines

Compared GWRH to two baselines that are analogous to state-of-the-art in chat clients:

- baseline_s – Check if *w* is a substring within a message
- baseline_t – Check if *w* is a match to a token within a message

Results

Highlighting Algorithm	Precision	Recall	F_2 score
baseline_s	0.71	0.52	0.55
baseline_t	1.0	0.5	0.56
GWRH	0.49	0.65	0.61

Results – Sanity Check

Top 10 relevant words:

- gnome
- classic
- 2d
- 3d
- compiz
- interface
- launcher
- plugin
- bar
- desktop

Conclusions

Designed a graph-based, unsupervised approach which improved upon recall when compared to current state-of-the-art.

Future work:

- Apply thread disentanglement (and address pronoun disambiguation)
- Life-time learning
- Human subject studies

Thank you!

References I

Berube, C. D., Hitzeman, J. M., Holland, R. J., Anapol, R. L., & 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., & Yilmazel, O. (2009). Recognizing connotative meaning in military chat communications. In *Proceedings of Evolutionary and Bio-Inspired Computation: Theory and Applications III*. SPIE.

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

Uthus, D. C., & Aha, D. W. (2013). The Ubuntu Chat Corpus for multiparticipant chat analysis. In *Proceedings of the AAAI Spring Symposium on Analyzing Microtext*, pp. 99–102. AAAI.