THE NPS CHAT CORPUS

- ➢ Corpus consisting of online chatroom conversations
- ➢ Part of the Natural Language Toolkit (NLTK) distribution.
- ➢ Tagged with Penn Treebank P.O.S. tags
- ➢ Available for download at Linguistics Data Consortium for $150.00 https://catalog.ldc.upenn.edu/LDC2010T05
- ➢ Produced by Eric Forsyth, Jane Lin & Craig Martell

OVERVIEW OF CREATION

- Compiled by researchers at the Department of Computer Science, Naval Postgraduate School
- Consists of 10,567 English posts gathered from various online chat services in October and November 2006.
- Each file is a text recording from one of these chat rooms for a short period on a particular day
- Work in progress, future versions will include more out of the approximately 500,000 posts collected

METADATA/ANNOTATION

Metadata

- Date, Target age group, Number of posts, Username
- Example: 10-19-20s_706posts.xml (19th October, 20s age group, 706 posts)

Annotation

- Each user given a generic name, UserN
- Posts were firstly tagged using a part-of-speech tagger trained on the Penn Treebank corpora
- Dialog-act tagging on the remaining posts was later done (e.g. <Post class="Statement/Question/System")

USE OF THE CORPUS

- Used to develop NLP applications that perform tasks such as conversation thread topic detection, author profiling, entity identification, and social network analysis.
- Used to help create and test micro-text classification methods in military chat (Rosa & Ellen, 2009)
- Used as a corpus to build an age group detection program made to analyse suspicious chat behaviour (Tam & Martell, 2009)

REFERENCES

- Forsyth, E. N., & Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. *Semantic Computing, 2007. ICSC 2007. International Conference on* (pp. 19-26). IEEE.
- Rosa, K. D., & Ellen, J. (2009). Text classification methodologies applied to micro-text in military chat. *Machine Learning and Applications, 2009. ICMLA'09. International Conference on* (pp. 710-714). IEEE.
- Tam, J., & Martell, C. H. (2009). Age detection in chat. *Semantic Computing, 2009. ICSC'09. IEEE International Conference on* (pp. 33-39). IEEE.
- The NPS Chat Corpus. (2007). Retrieved February 4, 2015, from http://faculty.nps.edu/cmartell/NPSChat.htm

ISLRN

Title	The NPS Chat Corpus
Full Title	NPS Internet Chatroom Conversations, Release 1.0
Resource Type	Primary Text
Source/URL	https://catalog.ldc.upenn.edu/LDC2010T05
Format	text/xml
Size/Duration	10,567 posts
Access Medium	Distribution: Web Download
Description	OLAC identifier: oai:www.ldc.upenn.edu:LDC2010T05 Release type: General Non-member fee: 150.00 USD Reduced-license fee: 150.00 USD Extra-copy fee: 150.00 USD Online documentation: http://catalog.ldc.upenn.edu/docs/LDC2010T05 Version 1.0
Media Type	Text
Language	English
Resource Creator	Eric Forsyth, Jane Lin, Craig Martell
Distributor	Linguistic Data Consortium
Rights Holder	Eric Forsyth, Jane Lin, Craig Martell