Filtering COMPS Chat Transcripts for Computer Modeling using Common Vocabulary

Nathaniel Bouman
Dept. of Computing and Information Sciences, Valparaiso University

Abstract

The Computer Mediated Problem Solving (COMPS) project aims to create a web-delivered problem-solving environment for student collaborative learning. One feature will be real-time computer monitoring of chat dialogs, informing instructors of the status and productivity of student discussions.

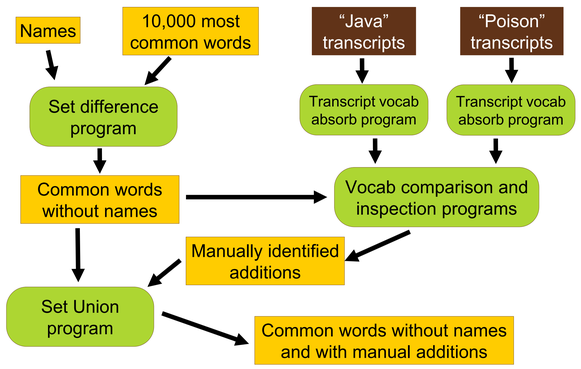
This work addresses challenges in preparing typed-chat of a variety of student exercises for computer analysis. Computer identification of productive chat will utilize a vocabulary of common English words not related to specific student exercises. Here we report on software and data resources for maintaining lexicons harvested from chat transcripts. This software aids in discovering vocabulary common to diverse chatting milieus, making the vocabularies available for research and for machine processing of the text. Typed chat also presents lexical challenges. Among them are words stretched looooonger or *starred* for emphasis, along with rampant spelling errors and abbreviations. Algorithms developed for these issues are presented here.

Where does this work fit in? This work in cooperation with L. Arndt, E. Graham, and S. Kalafatis, *Identifying Sudent Discussion in Computer-Mediated Problem-Solving Chat*, presented in this symposium. Their classifiers required:

- build/update a common vocabulary
- inspect transcripts
- filter text for issues with the typed lexicon
- remove or specially handle common vocabulary

Vocabulary Basis

- Started with list of 10,000 most common English words[1]
  - Google corpus built from 1 trillion words in text from public webpages[2]
  - > 90% everyday English usage coverage
- Needed to remove common person names from list
  - Set difference with 1990 US census data[3]
  - 90% US population name coverage.
  - *Most common* names don't vary very much over time
- Transcripts examined for words to add back manually
  - Slang and abbreviations.
  - Names that are more often regular words
  - Used programs described in "Vocabulary Inspection" and shown in diagram below



What Filtering Programs Do

Original Text	Text after Word Filters	Text after Word and Vocab Filters
This document is for testing filters	this document is for testing filters	this document is for testing filters
suchhh as stretching lettttterssss	such as stretching letters	such as $g01 letters
using *stars* for *emphasis	using $g02 stars for $g02 emphasis	using $g02 stars for $g02 emphasis
or, *the same thing* with multiple words	or $g02 the same thing with multiple words	or $g02 the same thing with multiple words
#and #thehashtag @at and ^carets	$g03 and $g03 $g04 and $g05	$g03 and $g03 $g04 and $g05
spellin erroprs of variouz types	spelling errors of various types	spelling errors of various types
and finally, junk: ixmvbulzew	and finally junk ixmvbulzew	and finally junk $g01

- Two types of filters: word and vocabulary. Applied in that order.
- **Word filters** perform operations individual words, based on lexical challenges discovered during vocabulary inspection.
  - **Filters for special characters** add a unique token identifying the character used.
    - $g02 for *, $g03 for #, $g04 for @, $g05 for ^.
    - For *, ^, and #, attached word kept if on common list
    - For *, only one token added if * surrounded the *start and end* of words.
  - **Filters for stretched letters** collapse the stretched letters to a word on the common list when possible.
    - No English non-hyphenated words have more than double letters.
    - All cases of duplicate letters changed to double letters.
    - All combinations of double/single letters checked: longest resultant word on common list (closest to original) chosen.
  - **Spelling filter** adapted from Peter Norvig's basic spelling checker[4].
    - All words 1 edit away from misspelled word checked.
    - Original Google 10,000 list is in frequency order: Choose resulting word with highest frequency.
    - Don't check words with non-alphabetic characters, or shorter than 4 characters.
- The word filters are important because the subsequent **vocabulary filter** replaces remaining words not on the common list with unique token $g01 (or removes the words, optionally).
- If words were not filtered first, valuable information could be lost through the vocabulary filter.

Vocabulary Inspection

Word	"Poison" Total Occurrences	"Poison" Transcript Appearances	"Java" Total Occurrences	"Java" Transcript Appearances	On Common Word List?
we	912	25	543	55	1
think	279	25	444	54	1
were	62	19	53	30	1
asking	3	3	36	22	1
skipping	0	0	3	3	0
wayyy	1	1	0	0	0
won	56	18	1	1	1
totalmortgage	0	0	1	1	0
encapsulate	0	0	5	4	0
*instantiating	0	0	1	1	0
#tired	1	1	0	0	0

Common words which occurred frequently

Common word not already on list to be manually added

Example of stretched letters phenomena

Problem-specific word considered common

Problem specific words *not* to be added to common word list.

Examples of special characters phenomena

- Additions to the 10,000 most common word list from Google needed to be identified through transcript examination.
- Programs extracted text from COMPS chat dialogs and compiled information to compare word usage between problem types.
  - Compared "Java" problem and "Poison" problem transcripts.
  - Total occurrences is how often a word occurred in total.
  - Transcript appearances is how many transcripts a word appeared in.
  - "On Common Word List?" marked 1 if "yes", and 0 if "no".
  - Research students used programs to identify and add words to common words list.
- Chart shown above has selected data.

Continuing Challenges

Although the filters catch many of the phenomena in the text, there are still some challenges to be addressed:

- **Some of the special characters have other uses**. (# as shorthand for "number", * to denote spelling correction, usage of all special characters to start an emoji, etc.)
- **Better decision making for the spelling filter**. Relative probability for words would be ideal. Other factors, (where the edit was made, what type of edit, etc.) could be used in combination with the probability.
- The words not originating from the Google 10,000 list have no probability or ranking attached to them.
- **Problem-specific words** that are one edit away from a word in the common vocabulary **are being corrected when we would like them not to be** (i.e. "getters" is changed to "letters" but should not be).
- The stretched letters filter could benefit from using word probabilities for cases with multiple possible corrections.
- The list of common words will need to be constantly updated with language trends.

Future Work

- Continued use of the vocabulary inspection programs and text filtering programs in concurrent work of computer modeling of COMPS transcripts.
- Research into problems listed in "Continuing Challenges" section above, particularly improvement of the spelling filter, as it is both the most complex and has the most room for improvement.
- Creation of versions of the filtering algorithms to work with the COMPS program in real-time to filter text for real-time versions of computer modeling programs.

Acknowledgements

- Professors Michael Glass and Melissa Desjarlais
- Fellow research students Lindsey Arndt, Emily Graham, Stamatina Kalafatis, and Yesukhei Jagvaral.

References (all accessed 27 July, 2017):

1. *google-10000-english.* Initial commit 29 Mar. 2012. https://github.com/first20hours/google-10000-english
2. Franz, Alex and Thorsten Brants. *All Our N-gram are Belong to You*. 3 Aug. 2006. https://research.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html
3. Meranda, Deron. *Names of People*. http://deron.meranda.us/data/
4. Norvig, Peter. *How to Write a Spelling Corrector*. Feb. 2007 to Aug. 2016. http://norvig.com/spell-correct.html

Partial support for this work was provided by the National Science Foundation's Improving Undergraduate STEM Education (IUSE) program under Award No. 1504918. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.