A Review of NLP Annotation tools and Platforms for Semantics and Pragmatics

By
Saptarashmi Bandyopadhyay
First Year PhD Student
Department of Computer Science and Engineering
Pennsylvania State University, University Park

Outline

- Overview
- Research Motivation
- Collection of Annotation Data
- Usability Measures Proposed
- Recommendations of User Interface of the Tools
- Task Analysis
- Future Work

Overview

- Natural language annotation: Metadata tags to mark up certain sections of texts in a dataset [1]
- Annotation should be accurate and relevant to the task [1]
- Important for smart HLTs (Human Language Technologies) [1]

Research Motivation

- Importance of natural language annotation
  - analysis of dialog systems
  - quality of summarization
  - analysis of data-sets generated by the social media and several other applications
- Currently few literature resources available on NLP annotation tools [2]
- Review several NLP annotation tools and platforms in
  - Semantics to find meaning from the text data
  - Pragmatics which refers to language in use depending on the context

Current Literature Overview [2]

- Author: Dr. Manuel Burghardt,Head of Computational Humanities group, University of Leipzig
- PhD dissertation on the same topic: Engineering Annotation Usability - Toward Usability Patterns for Linguistic Annotation Tools
- 13 citations till date in Google Scholar

Current Literature Contributions[2]

- Heuristic usability evaluation of three annotation tools
  - GATE [3]
  - MMAX2
  - UAM Corpus-Tool
- General problems originating from ignoring established best practices and guidelines for user interface (UI) design,
- Specific problems which are domain-dependent on linguistic annotation.
- Developing awareness among tool developers.
- Set of 28 design recommendations
  - Describing generic solutions for the identified problems
  - Structured and systematic collection of usability patterns for linguistic annotation tools.

My Contributions

- Review of NLP annotation tools
- User interface recommendations
- Identification of potential drop-out users
- Development of generic usability metrics
- Strategies and behaviors of users
- Identification of important sub-task during annotation

Annotation Tools

- GATE [3] is a general purpose NLP text engineering tool which users find cumbersome for manual annotation
- DucView [4] annotation tool
  - Creation of pyramid model [7] of content units and their importance
  - Derives importance weights for content units from multiple human summaries

Perspectives

- Usability of the annotator: The goal is to identify usability issues of the annotation tools when used by the annotator.
- Usability of the researcher: The goal is to simplify researcher’s use of the annotation tools while training the annotators.

Data

- User interviews from verbal protocol analysis [8]
- Users from ongoing annotation project in NLP Lab, Pennsylvania State University, under Prof. Rebecca J. Passonneau

Collection of data

- Interview of user for estimated 6 or 7 minutes for using the annotation tool on a very small paragraph provided
- Use of the tool in an estimated 4 or 5 minutes,
- Summary of the experience in an estimated 2 minutes to generate the retrospective report
- Audio recording

Recording medium

- 3 participants (User 1 and 2 of DucView and User 3 of GATE) have been interviewed on-campus which has been recorded with due permission by the Sound Recorder application in my smart phone.
- 3 participants (User 1 and 2 of GATE and User 3 of DucView) have been interviewed on Skype with the call being recorded with due permission.

Principle of data collection

- Talk-aloud protocol [8]
- While using the annotation tool, user has to talk about what he/she would have said to himself/herself
- User has to keep talking
- Prodding to avoid any pause in recording
- Retrospective report
- Summary of what the user remembers of his thought process at the end of the task

Purpose of Retrospective Report

- Talk-aloud protocol gives an overview of the task being conducted
- Still retrospective report being collected to understand the short-term memory and the long-term memory of the user (their cognitive capability)
- Useful to evaluate the biasness in selection of users
- Ensures completeness of the talk-aloud protocol

IRB guidelines

- Willing consent of the user for the study
- Benefit of the user
- Protection of the privacy of the user
- User cannot be identifiable from the metadata

Permission from IRB

- Class project so not required
- Permission needed for funded research projects or publications

Users for GATE

- 15 annotators working in the Annotation project at NLP Lab, PSU, out of initial 50 annotator responses.
- 3 out of 15 participated in the usability annotation project
- Useful in improving the quality of the Annotation project at NLP Lab, PSU
- Online user reviews (non-reliable data)

Users for DucView

- 2 first-time users of DucView
- 1 experienced user of DucView
- Useful in contrasting the role of experience in strategy selection for task analysis

Warm-up activity

- User was given a quiz to understand intent to participate
- 32 quizzes (data obtained for the class project with permission from the researchers in the ongoing Annotation project at NLP Lab, PSU)
- A tutorial to download the tool and understand its uses

Snapshot of the Quiz

Read the following text and answer the question

`` When writing a novel a writer should create living people; people not characters. A character is a caricature." Ernest Hemmingway. Characters are essentially what make and break the differences between the novella and the movie. In the novella, the father and mother are so believable, especially the fathers reaction to having an elderly man as a son, and that is portrayed by how he treated the tailor, and the hospital staff.

Is "that" referring to an object, or a fact/event/idea/concept/action …?

Could you find the raw text "that" refers to?

Identification of potential drop-outs

- Users who stopped giving their feedback in the middle of the study (dropped out) have similar completion times of the warm-up task
- If same quiz is used in future, then potential drop-outs can be identified
- Otherwise the dataset becomes useless when user drops in the middle of a task
  - Such data can be used to act as dummy users

Overview of drop-out identification

- Annotators who took less time (8-10 minutes) on the quiz continued in the project
- Annotators who took more time (>10 minutes) on the quiz dropped out.
- Threshold time of solving the quiz (10 minutes) can be identified to prevent user drop-outs in an ongoing task.

GATE User Data

User	Time taken for talk aloud protocol	Words in the transcript in talk aloud protocol	Time taken for retrospective report	Words in the transcript in retrospective report
1	6 mins 26 seconds	329	33 seconds	50
2	5 mins 09 seconds	700	1 min 33 seconds	246
3	6 mins 08 seconds	598	1 min 50 seconds	275

DucView User Data

User	Time taken for talk aloud protocol	Words in the transcript in talk aloud protocol	Time taken for retrospective report	Words in the transcript in retrospective report
1 (amateur)	6 mins 27 seconds	457	1 min 11 seconds	107
2 (amateur)	6 mins 39 seconds	330	1 min	81
3 (experienced)	5 mins 36 seconds	762	1 min 8 seconds	159

Usability Metrics

- Current ISO standards usability metrics are dependent on the user
  - Effectiveness
  - Satisfaction
  - Productivity
  - 7 other factors [9]
- Task Effectiveness (T.A.) = (Quality * quantity)/100
- Quality is the proportion of the goal achieved
- Definition of the goal can be subjective

Factors influencing the usability metrics

- Fair selection of users
- Cognitive capability of users
- Focus on the task by the users
- Bias of the experimenter
- Reliability of the dataset
- Impact of users on other users
- Motivation of users
- Weightage of the generic issues

Measure of Cognitive capability of user

- Can be measured based by the number of words in talk-aloud protocol divided by the duration of the talk-aloud protocol
- For user 3 of DucView, the ratio is 762/5.6 = 136.071428571
- For user 2 of DucView, the ratio is 330/7.15=46.15384615384615
- User 3 is actually more experienced than user 2 and the transcript data is more useful for analysis

Use of the Measure

- It can be used also for measuring the motivation of the user to find which transcript should be given more importance.

Impact of One User on the Other User

- Bag of words which are apologetic or dithering like ‘sorry, umm’
- User 1 of DucView says sorry 9 times out of 457 words in talk aloud protocol

Focus on the Task

- The focus on the task can be measured by the weightage of the off-topic words (kept in a bag of words) from the total number of words in the transcript.

Fair Selection of Users

- The summary (retrospective report) can be compared to the transcript of the talk aloud protocol by ROUGE score
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score for evaluating the quality of the summary
- Similar ROUGE scores indicate, that the short term and long term memory of the users is similar which denotes that the selection of users was biased and the bias can be considered by an offset factor.

Bias of the Experimenter

- Subjective Evaluation
- Introduced while removing noise while transcribing the audio recordings
- Bias due to perception of noise
- Bias due to domain knowledge of the researcher

Recommendations for DucView and GATE Tools

- Controls of the software could be highlighted better with a better GUI(graphical user interface).
- Selection of particular files is not highlighted clearly which indicate the need to provide footnotes with each function.
- Confusion of the subjects illustrate more explicit instructions in the user manual.
- Processing of the software is very slow in GATE which might give some visualization of the task or might indicate the intellectual prowess of the subjects working on the task.
- Context in which the usability experiment is being conducted is important.

User Behavior

- Interesting psychology of repeating different words confidently and repeatedly which the speaker does not mean to say actually and is unaware of
- One user mumbled while reading some portions of the text despite being prodded repeatedly to talk aloud for those 14 seconds
- While talking about the task, one user is taking actions and trying to reason the actions to own self.

Task Analysis:-Introduction

- Useful in describing and understanding how a task is to be performed.
- Description of the user's tasks at different level of abstractions [10]
  - Hierarchical Task Analysis (HTA) where goals are decomposed to sub-goals with visual representation
  - Cognitive Task Analysis (CTA) describing how all tasks are completed.
  - GOMS (Goals, Operations, Methods and Selection rules) specifying the error-free, expert behavior, multiple strategies to solve the same task.
  - Keystroke Level Model (KLM) used to compute the time taken for a single task completion by the users.
- Profound applications in improving the performance of army systems and savings of billions of dollars[11] due to it's application in human systems integration (HSI).

Task analysis

- The most important sub-task of using the annotation tools can be identified by task analysis
- Saving significant investment in training the annotators to use such tools
- Hierarchical task analysis

Task analysis on Using NLP Annotation Tools

Task: Usage of NLP annotation tools

No.	Sub-task
1	Open the tool
1.1	Movement of mouse to Linux terminal icon
1.2	Click the icon
1.3	Typing of the command to open the tool
1.4	Pressing Enter in keyboard
2	Open the file button
2.1	Locate the file button
2.2	Movement of mouse to the file button
2.3	Click on the file button
3	Opening the file
3.1	Click on the new button
3.2	Move the mouse to search the relevant file
3.3	Click on the desired file
3.4	Move the mouse to the open button
3.5	Click on the open button
4	Annotation on the file
4.1	Moving the mouse to the particular phrase
4.2	Selection of the phrase
4.3	Move the mouse to the new annotation unit button
4.4	Click new annotation unit button
5	Change the annotation label
5.1	Move the mouse to the label
5.2	Click on the label
5.3	Move the mouse to the change label button
5.4	Click the change label button
5.5	Move the mouse to the text box
5.6	Type the new name from the keyboard
5.7	Move the mouse to the OK button
5.8	Click OK button
6	Backup of the annotation file
6.1	Click Ctrl-S keys in the keyboard

Task Analysis on Using NLP Annotation Tools (contd)

7	Exiting the annotation task
7.1	Move the mouse to file icon
7.2	Click the file icon
7.3	Move the mouse to Save As icon
7.4	Click the Save As icon
7.5	Type the desired name of the file to be saved from the keyboard
7.6	Move the mouse to the Save As button
7.7	Click the Save As button to terminate

Effect of Domain Knowledge on Task Analysis

- Users were told to complete annotation on DucView or GATE tool.
- They were not told any strategy.
- Experienced users applied strategies like reading the document line by line to find SCUs in DucView which led to better annotation.
- Study of the strategies required for better learning

Alternative to Interview

- Eye tracker
- Key stroke tools like RUI
- Challenges of user availability

More tools and platforms

- Brat [5] is a web-based annotation tool to capture annotation of social media data
  - No current usability resource available
  - Context of annotation can be uncertain
  - Has been used for many kinds of annotation, including transactivity detection (Prof. Rose, NLP Colloquium talk)
- Amazon Mturk [6] is a platform for crowd sourcing where Human Intelligence Tasks (HITs) are given to workers.

Challenges of Amazon mturk – User Motivation

- Data from turkers are often not reliable as the turkers are not motivated enough
- Objective is often the reward
- Data is collected by employing annotators from the specified domain, so more interested.

Outline of Completion of the Project

- Calculation of ROUGE score to evaluate the quality of the summary in the retrospective report
- Completion of project report

Future Work

- Keystroke level model can be used to measure the task timing
- Context identification in the bag of words by similarity scores
- Automated transcription using speech to text tools like CMU Sphinx
- Think-aloud protocol to study the impact of emotions of user's behavior on the usability study

References

1. Natural Language Annotation for Machine Learning book by Amber Stubbs, James Pustejovsky, O'Reilly Media, Inc., October 2012, ISBN: 9781449332693
2. Usability recommendations for annotation tools by Manuel Burghardt, pages 104-112, Proceedings of the Sixth Linguistic Annotation Workshop (LAW VI '12), Association for Computational Linguistics, 2012.
3. Developing Language Processing Components with GATE Version 8 by Cunningham, et al., GATE User Guide, University of Sheffield Department of Computer Science, 17 November 2014.
4. http://www1.cs.columbia.edu/~becky/DUC2006/2006-pyramid-guidelines.html
5. http://brat.nlplab.org/
6. https://www.mturk.com/

References

7. Evaluating Content Selection in Summarization: The Pyramid Method by Ani Nenkova and Rebecca Passonneau, N04-1019, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL, 2004, http://www.aclweb.org/anthology/N04-1019

8. Usability measurement and metrics: A consolidated model by Ahmed Seffah, Mohammad Donyaee, Rex B. Kline, Harkirat K. Padda, Software Quality Journal, June 2006, Volume 14, Issue 2, pp 159-178, DOI: 10.1007/s11219-006-7600-8

9. Verbal Reports as Data by K. Andersen Ericsson and Herbert A. Simon, Psychological Review, Volume 87, No. 3, pages 215-251, 1980

References

10. Methodology I: Task Analysis by Frank E. Ritter, Gordon D. Baxter and Elizabeth F. Churchill, Chapter 11 of the book 'Foundations for Designing User-Centric Systems' published by Springer-Verlag London in 2014, page nos. 309-333, DOI: 10.1007/978-1-4471-5134-0_11.

11. Human Systems Integration in Army Systems Acquisition by Harold R. Booher and James Minniger, Chapter 18 of the book 'Handbook of Human Systems Integration' published by Wiley Online Library in 2005, page nos. 663-698, DOI: 10.1002/0471721174.ch18

THANK YOU