THE PENNSYLVANIA STATE UNIVERSITY

Project Title:

A Review of Natural Language Processing (NLP) Annotation Tools and Platforms for Semantics and Pragmatics

by

Saptarashmi Bandyopadhyay

szb754@psu.edu

December 11, 2018

First Year Ph.D student

Department of Computer Science and Engineering

Pennsylvania State University

University Park, PA 16802

Class Project for IST 521

Table of contents

Table of contents (continued)

1. ABSTRACT

The objective of the class project is to review two Natural Language Processing (NLP) annotation tools, GATE and DucView that carry out natural language annotation in the domain of semantics and pragmatics as there are very few review resources of usability for NLP annotation tools. Experienced as well as first-time users of GATE and DucView have been interviewed for the project using talk-aloud protocol and retrospective reports. A set of user interface recommendations have been proposed to improve the interface of GATE and DucView tools. A set of usability metrics have been proposed namely fair selection of users, cognitive capability of users, focus on the task by the users, impact of users on other users and motivation of users have been proposed along with a mechanism of objective evaluation. Other usability metrics identified like bias of the experimenter and reliability of the dataset is subjective in nature. These measures can be applied over all the users and as a result can give a generic evaluation of the usability of the interfaces, compared to current measures that are dependent on the users. The most important sub-task in the natural language annotation task has been identified by Hierarchical Task Analysis (HTA). A mechanism has been proposed to identify the users who can potentially drop out while using the annotation tool. Such users can be identified based on the completion time of their warm-up activity. It has been observed that the users are adopting different unique strategies during the talk-aloud protocol to use the NLP annotation tools which can be used to make the tools adapt to the strategies of the users.

2. INTRODUCTION

Natural language annotation involves identification of metadata tags to mark up certain sections of texts in a dataset [1]. Annotation should be accurate and relevant to the task [1]. Annotation of natural languages is important for smart Human Language Technologies (HLTs) [1]. Importance of natural language annotation is evident in the analysis of dialog systems, evaluating the quality of summarization, analyzing data-sets generated by the social media and several other applications. The domains in which the NLP annotation tools have been used are i) Semantics which involves finding meaning from the text data and ii) pragmatics which refers to language in use depending on the context.

Currently only one literature resource is available on the usability of NLP annotation tools by Dr. Manuel Burghardt, Head of Computational Humanities group, University of Leipzig which was also his PhD dissertation, “Engineering Annotation Usability - Toward Usability Patterns for Linguistic Annotation Tools” [2]. The paper on usability annotation has 13 citations till date in Google Scholar[2]. Heuristic usability evaluation has been carried out in that paper on three annotation tools GATE [3], MMAX2 and UAM Corpus-Tool. It highlights the general problems originating from ignoring established best practices and guidelines for user interface (UI) design and specific problems which are domain-dependent on linguistic annotation. The paper serves as a medium of awareness among tool developers. 28 design recommendations have been provided which describe generic solutions for the identified problems and involve structured and systematic collection of usability patterns for linguistic annotation tools.

The data has been collected using talk aloud protocol and retrospective reports based on the idea that verbal reports can be treated as data. It is to be demonstrated that accounting of verbal reports require detailed explanation of the mechanisms, used to generate the reports and the processes by which these reports are sensitive to experimental factors just like in case of other kinds of data. Verbalization affects cognitive processes only if the instructions require verbalization of information that would not have been attended to otherwise and that inaccuracies in the reports are due to inferences by the subjects. These ideas have been presented in the paper[9],'Verbal Reports as Data' , as well as in the appendix of the book 'Protocol Analysis'. The paper and the book are written by by Prof. K. Andersen Ericsson, who was a post-doctoral research scholar at Carnegie Mellon University at that time, and currently Conradi Eminent Scholar and Professor of Psychology at Florida State University. They are co-authored by noted scientist Prof. Herbert A. Simon, affiliated to Carnegie Mellon University and being the recipients of the Turing Award in 1975, Nobel Prize in Economics in 1978, National Medal of Science in 1986, A.C.M Fellow in 1994 among other distinguished accomplishments.

Task analysis (TA) can be used to find the most important sub-task in a task. It is useful in describing and understanding how a task is to be performed. Some widely used methods of TA to describe the user's tasks at different level of abstractions [10] are :
i) Hierarchical Task Analysis (HTA) where goals are decomposed to sub-goals with visual representation
ii) Cognitive Task Analysis (CTA) which describes how all tasks are completed.
iii) GOMS (Goals, Operations, Methods and Selection rules) which is a part of CTA but has an important role in HCI where specification of the error-free, expert behavior allows for multiple strategies to solve the same task.
iv) Keystroke Level Model (KLM) which is again a part of CTA but is considered separately due to its close relationship with HCI where it can be used to compute the time taken for a single task completion by the users.

The challenging task is to identify which among the different task analysis methods has to be applied for this particular circumstance to execute task analysis. A task analysis tool, Herbal[17] has been studied for the above purpose. The role of TA in risk-driven incremental commitment model and in general Human-Computer Interaction (HCI) has been discussed at the end of the laboratory report. A task has been selected which is “Using two Natural Language Processing (NLP) annotation tools” and all the sub-tasks are identified to ensure their total coverage before successful usage of the tools. The selection of task analysis methods is dependent on the risks involved like timing of completion of tasks, their interface design and interaction with the user, coverage of all sub-tasks and process coverage of all the tasks among other risks. Hierarchical Task Analysis method has been selected which has been justified in the Section 3.1.2.

The role of task analysis has profound applications e.g. it has improved the performance of army systems and saved a huge amount of money[16], [11] due to its application in human systems integration (HSI). Identifying the most important subtask during NLP annotation will allow the researchers to train the annotators for only that subtask, that will save the time for experiment and the salary expenditure for annotators.

3. METHOD

3.1 Task

A task of using two NLP annotation tools was identified to carry out task analysis and to divide the task into sub-tasks with the help of operators. The two NLP annotation tools selected for the experiment were GATE [4] and DucView[5],[6].

The DucView tool, executing pyramid annotation, was developed by Dr. Ani Nenkova, who was at the time, a Ph.D student in Computer Science Department, Columbia University, New York, U.S.A. and currently Associate Professor, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, U.S.A. and Prof. Rebecca Passonneau, previously affiliated to the Computer Science Department, Columbia University, New York, U.S.A. and currently affiliated to the Department of Computer Science and Engineering, Pennsylvania State University, University Park, U.S.A. A verbal protocol analysis[7] applying the talk-aloud protocol, as described in the appendix[8] of the book 'Protocol Analysis', was applied to three of the six users in a study room and from remote location by Skype who had used the DucView tool in the previous Verbal Protocol Analysis laboratory and were thus experienced with the DucView tool. An user was told to keep talking about what the user would have said to himself silently while using the tool which was recorded and transcribed. The talk aloud protocol was used instead of the think-aloud protocol in order to preserve the temporal properties of the sub-tasks to understand the mechanism of cognitive processes and decomposition of the tasks into sub-tasks.

The GATE tool[4] is a general purpose NLP annotation tool which has been developed by a 15 member research team led by Prof. Hamish Cunningham, Department of Compuer Science, University of Sheffield, U.K. Some of the user data for GATE tool has been collected from online discussion of users on using the GATE tool from online platforms of discussion like ResearchGate[11], Quora[10] and StackOverflow[9] due to paucity of users who could be interviewed for a laboratory experiment. The users are discussing about GATE version 8. The dates of discussion are after the release of version 8 which require JAVA 8 or higher SDK (software developer's kit) in any laptop or P.C.

The DucView tool, version 1.4, was reviewed by the two users (A and B) in a laptop with Linux Operating System, Fedora 20 and by one user (C) in a laptop with Mac High Sierra 10.13.6 Operating System. The manual was given to them to read about the tool before the warm-task was carried out for one minute giving the users a small task of using DucView on a small paragraph for annotation to make them comfortable in talking aloud. Then the verbalization for the main task was recorded in a sound recorder application of the mobile phone ore recorded on Skype, based on the method of interaction with no external noise interference for six minutes which was then manually transcribed for analysis. Then the users were requested to remember what they said in the talk-aloud protocol which was recorded as the retrospective report. This recording was done for approximately two minutes. The same text was annotated by the 3 users.

The GATE tool was reviewed by the one user (C) in a laptop with Linux Operating System, Fedora 20 and by two users (A and B) in a laptop with Mac High Sierra 10.13.4 Operating System and Mac High

Sierra 10.13.6 Operating System. All of their laptops had JAVA Version 1.8. The users were provided a manual of how to download and use the GATE tool. Warm-up activities as quizzes to apply co-specification annotation in example problems were provided to the users by researchers at the NLP Lab, P.S.U. The main task was verbalized and recorded in a sound recorder application of the mobile phone ore recorded on Skype, based on the method of interaction with no external noise interference for six minutes which was then manually transcribed for analysis. Then the users were requested to remember what they said in the talk-aloud protocol which was recorded as the retrospective report. This recording was done for approximately two minutes. The same text was annotated by the 3 users.

From the analysis of the above data, TA has been carried out on using the 2 annotation tools. The TA has been represented in Herbal[17] which is a high level language for behavior representation that can be used to develop and support the variety of users of intelligent agents and cognitive models. The Herbal tool has been developed by Prof. Frank E. Ritter, Director of Applied Cognitive Science Lab, College of Information Science and Technology, Pennsylvania State University, University Park, U.S.A, Dr. Jong W. Kim, who was a Ph.D student in the College of Information Science and Technology, Pennsylvania State University, University Park, U.S.A and currently a Post-doctoral Research Associate at the University of Central Florida, Dr. Mark A. Cohen who was a faculty member at Lock Haven University, Lock Haven, U.S.A. and is currently affiliated to the Department of Computer Science at MCLA (Massachusetts College of Liberal Arts) and Prof. Steven R. Haynes, from the College of Information Science and Technology, Pennsylvania State University, University Park, U.S.A. Herbal tool of version 3.0.5 has been downloaded in a laptop with Operating System Fedora 20 and Eclipse version of 4.3.2 which is compatible with Herbal tool. The downloaded jar file is saved in the drop-in folder of Eclipse which is then opened to used the Herbal tool.

3.1.1 Reason for selection of Herbal Tool for Task Analysis (TA)

The user behavior and interaction with the annotation tools needs to be modeled. Herbal tool facilitates the modeling of the user behavior to understand the cognitive processes or sub-tasks needed to complete the task. As Herbal operates as an Eclipse drop-in on JAVA, it provides explainable code which helps in clear understanding of the decomposition of sub-tasks required to use the two annotation tools. It supports the representation of the task as the intelligent agent, the sub-tasks as the problem spaces and the operators are defined as the steps required to go from one sub-task to the other. Thus it is useful in representation of the Hierarchical Task Analysis (HTA) method on the task of using the annotation tools. The reason for selection of the HTA method has been described in the Section 3.1.2.

3.1.2 Reason for selection of HTA for Task Analysis

The Hierarchical Task Analysis (HTA) has been carried out for task analysis of using the two NLP annotation tools. This is because selection of the task analysis method is dependent on the risks involved in design of the system. Here, the objective was to observe, which tasks are necessary to annotate in both the tools and whether the tasks are general or specific in the two tools whose functions are different from one another. HTA is a structured process to record the user performance while carrying out a task. It is highly useful to understand the complete coverage of all the sub-tasks to fulfill an aim.

This approach is also used in optimization of the steps required to go from one sub-task to the other. Often, the tasks and its decomposition into multiple sub-goals can be similar but there can be multiple strategies in order to execute one sub-task from another. Often the sub-tasks are different but the end objective is same due to use of different problem approaches. HTA will help in analysis of the multiple strategies and identification of the best strategy required to decompose the task into many sub-goals. HTA provides a generic representation to the different strategies that can lead to the completion of a goal.

The different strategies can lead to re-ordering of the sub-tasks in order to ensure better performance in execution of the work at hand[10]. If some sub-tasks have no impact from some other sub-tasks, these set of tasks can be executed in parallel as they are independent of each other. This leads to optimization of the task objective and consequent performance improvement. The analysis of the impact of cued action on system performance is reflected by HTA.

The precise details of steps for execution of the tasks and sub-tasks can lead to efficient development of systems if the information is used appropriately by the engineers. Since, the interface is abstracted in this task analysis method, it is very useful in reflecting the early system development[10]. It also helps in understanding how reuse of the user experience design components can be carried out based on the sub-tasks which are repeated continuously during execution. The HTA can be extended to reflect he cognitive processing of the users and the tasks regarding how the task coverage is achieved. As the focus is on the details of the task, a talk-aloud protocol has been applied as a verbal protocol in collection of data instead of the talk-aloud protocol.

Based on the analysis of the data in Section 4.1 of Results, the keystroke level model could also have been used in addition to the Hierarchical Task Analysis. Both the users hint at problems in the interface of the tools which is not adequately represented in the hierarchical task analysis. Further data of keystrokes could have been collected with tools like RUI (Recording User Input) in order to conclude the effect of the interface in completion of the task. However, that would have been complicated as the users did not consent to their recording of their signature of their writing speeds and their user inputs. The analysis would also have been complex as it would have been in addition to the current analysis for a laboratory assignment. However, extension to KLM can be considered as future work.

3.1.3 Role of TA in RD-ICM Model and HCI

Task analysis has a major role in RD-ICM (risk driven-incremental commitment model). The tasks, whose performance is mandatory can be labeled as to-do tasks[21] which helps in saving costs in each life-cycle of the software development by removing such tasks from the incremental model. Such tasks can be identified by carrying task analysis and from the measurement of the number of steps of the sub-tasks or the number of sub-tasks of the task.

Task analysis ensures that the actions of the users are fulfilling the task objective [22]. Also the user behavior can be predicted in HCI tasks from the data obtained during task analysis and by proper information sharing among the users [22]. The tasks can be mapped to the behavior of the user [22].

3.1.4 Purpose of Retrospective Reports

Talk-aloud protocol already gives an overview of the task being conducted. Still the retrospective reports are being collected to understand the short-term memory and the long-term memory of the user which gives an idea of their cognitive capability. The retrospective reports can be considered as summaries of the talk aloud protocol as while the user remember retrospectively the talk-aloud protocol, he or she remembers the primary issues which is basically a summary. Thus it can be considered that the talk-aloud protocol is a manifestation of the short-term memory of the users while the retrospective reports can give an idea about the long-term memory of users

Retrospective reports are useful to evaluate the biasness in selection of users. For one user, the summary generated from the retrospective reports can be used as reference to evaluate the quality of the summaries of the other users. Lower and different values of precision, recall and F measures indicate that the other summaries are different compared to the summary of the user. This means that the cognitive capability of the users to summarize is different and it can be declared from the objective scores that the selection of the users was unbiased. However, higher and similar scores of precision, recall and F measures indicate that the selection of users was biased.

Another reason behind combining the talk-aloud protocol and retrospective reports to a common warm-up procedure is to ensure completeness of information. It is expected that for cognitive processes of intermediate duration, it is expected that both of them should have similar information. Hence, if the subject gives both reports for the same cognitive process, it would help to assess the completeness of the talk-aloud protocol and that retrospective report contains an actual record of the cognitive process.

3.2 Subjects

Six subjects have been used for the purpose of the experiment. Three subjects were interviewed for the DucView tool of whom two were first-time users and one had previously used DucView. Three subjects were interviewed for the GATE tool. Based on the guidelines mentioned in the Belmont report under the regulations of Institutional Review Board (IRB) for human subjects research, the subjects gave their consent willingly to be a part of the project which would benefit their understanding of task analysis in their respective research areas. No prejudice was exerted in the selection of the six users as per the federal regulations of the National Research Act. The subjects were assured that they will not be harmed in the experiment and their data will remain private.

The experienced users of the NLP annotation tools had willingly taken part in the online discussion of the usable annotation tools in public platforms[18], [19], [20] which manifests their intention to share the information to everyone for the greater good of research. Their participation in the public discussion is manifestation of their consent and willingness to benefit from being a part of the learning process and thus IRB guidelines were followed in using the data provided by the users in online platforms on the annotation tools.

The laboratory assignment was exempt from IRB review as normal educational practices were used in the class project as per the Common Rule, 45 CFR 46.101.

3.3 Design and Procedure

3 users A, B and C were interviewed for using the DucView tool. Users A and B were interviewed face-to-face in a study room in order to study the impact of the talk-aloud protocol on the first user on the talk-aloud protocol executed by the second-user. and the recordings were taken in the sound recorder application of an Asus cell phone. User C was remotely interviewed and recorded on Skype. The recordings were initiated with permission of the users.

3 users A, B and C were interviewed for using the GATE tool. Users A and B were remotely interviewed and recorded on Skype. User C was interviewed face-to-face in a study room and the recordings were taken in the sound recorder application of an Asus cell phone. The recordings were initiated with permission of the users.

The user interview by talk-aloud protocol was conducted for approximately 6 minutes while the user interview by retrospective report was conducted for approximately 2 minutes. However, individual users finished the task a bit earlier or later than the stipulated time as shown in Section 4.1.

All 6 users for the 2 natural language annotation tools had to keep talking continuously while using the tool what they would have told to their ownselves while in a room all alone. They were requested not to plan or explain their thoughts but to tell it as it was coming in their thought process. It was informed to them that they could refer to the manual anytime while using the tools and if they became silent, they would be prodded to talk.

Initially warm-up activity had been conducted by providing quizzes of the co-specification annotation task to user of the GATE tool by researchers at the NLP Lab, Pennsylvania State University (PSU). The users of the GATE tool were paid for the duration of the interview as a part of an ongoing research project at the NLP Lab, PSU. The users of DucView volunteered to use the tool without any salary.

A small task of one minute based on a file of three lines was given to each one of them for annotation to warm-up with the talk aloud protocol so that they did not feel shy in talking aloud in front of me or the other user. Then the primary task was recorded for approximately six minutes in the sound recorder application of my cell phone or Skype for each of the users depending on the mode of interview. The recordings for the users were heard multiple times for manual transcription by ignoring noises or any incoherent section of the recording,

The online discussions in three public discussion platforms of StackOverflow[9], Quora[10] and ResearchGate[11] regarding using of NLP annotation tools were also collected as data for the task of using the GATE annotation tool. It had been assumed that the users are sharing their experience with the GATE tool correctly based on the use of the above mentioned freely accessible public platforms in a large scale for research discussions. There were six answers in the Quora and StackOverflow platforms each while there were three answers from users in the ResearchGate platform. The same user could have given multiple feedback with multiple usernames but as each of the answers were unique based on the policies of such public platforms, it would not matter in the final analysis.

The data obtained by the above two techniques has been analyzed and accordingly verbal protocol analysis and task analysis has been carried out which has been shared in the Section 4 of Results.

The herbal tool has been used for representation of the task analysis. The tutorial was executed to create the intelligent agent, its problem space and set of operators for a hungry and thirsty agent called Sally which has a single problem space of survival for which she has two parameters of 'eat' and 'drink' in order to survive. Based on that training, the herbal tool has been used to represent the task analysis in this laboratory assignment. The task has been used as the agent while the sub-tasks are used to denote the problem-space. The steps carried out to complete one sub-task after another sub-task are considered as operators in modeling the user behavior in the Herbal tool.

3.4 Analysis of the method

3.4.1 Advantages of the talk-aloud protocol with retrospective reports

The advantages of the talk-aloud protocol with retrospective reports can be listed as following:

1) The retrospective report carried after execution of the talk-aloud protocol ensures completeness of the information, by covering any possible information that might be missing in the concurrent verbalization.

2) The talk-aloud protocol avoids ambiguities of specific question answering of yes or no, where the subjects might infer leading to incorrect reports.

3) The cognitive processes can be studied intricately as the instructions are requiring verbalization of all possible information in the thought process.

4) The process is extremely data rich and provides a large data-set for transcription even for recording for a small amount of time.

5) Details of intermediate processing can be studied if the temporal density of the observation can be increased during the protocol.

6) The warm-up procedure helps the first-time users to be comfortable with the environment of the experiment and how to perform the task properly.

3.4.2 Disadvantages of the talk-aloud protocol with retrospective reports

The advantages of the talk-aloud protocol with retrospective reports can be listed as following:

1) The verbal protocol is highly dependent on the user. There are many aspects of the user that are to be investigated like emotions.

2) The talk-aloud protocol does not provide complete visualization of the task as it does not provide focus on the task objective but only on the interaction of the subject with the task.

3) The researcher can be biased in encoding the transcription from the recording to incoherent noise as

a wrong judgment will lead to loss of valuable data.

4) The transcriptions are highly voluminous and appropriate tools need to be used to analyze and understand the results.

5) The focus on the task during concurrent activity by the subject is a considerable challenge as it depends on the intellectual capability of the subject.

6) The process of data collection is bounded by time, although it may not reflect the completion of task by the subject and thereby affecting the conclusion.

7) The possibility of a guided conclusion is not considered in the talk-aloud protocol which could have provided significant insight in the cognitive model of cooperation between the subject and the researcher.

3.4.3 Analysis of HTA

The HTA on the task of using NLP annotation tools is influenced by a number of factors which can be listed as follows:

1) The method is based on the selection of users which plays an important impact on the results. As the users are experienced in using both the tools, their cognitive capabilities improve the task analysis mechanism compared to the first-time users.

2) The experiment generates large amount of data from verbal protocol analysis even though the recording is done for only a small amount of time. Appropriate tools need to be used for the analysis of large quantities of data.

3) Details of the intermediate processing can be studied with respect to time.

4) The impact of the assumption that the users are speaking truthfully in their online discussion is significant on the results.

5) The impact of the user's talk-aloud protocol influenced by the other user in the room who is using DucView opens an interesting area on the efficiency of task analysis.

6) The experimenter can be biased in identifying which noise is incoherent or not, primarily if the experimenter and the subjects know each other before. Unfortunately, a wrong judgment can cause the loss of important data.

4. Results

4.1 Description of the Data

User	Time taken for talk aloud protocol	Words in the transcript in he talk-aloud protocol	Time taken for retrospective reports	Words in the transcript in retrospective report
A	6 mins and 27 seconds	457	1mins and 11 seconds	107
B	6 mins and 39 seconds	330	1 mins	81
C	5 mins and 36 seconds	762	1 mins and 8 seconds	159

Table 1: Data about users using the DucView tool

From Table 1, for user A, the recording was 6 minutes and 27 seconds in the talk-aloud protocol and 1 minute and 11 seconds for the retrospective report which has been transcribed to 457 words and 107 words respectively.

For user B, the recording was 6 minutes and 39 seconds in the talk-aloud protocol and 1 minute for the retrospective report which has been transcribed to 330 words and 81 words respectively.

For user C, the recording was 5 minutes and 36 seconds in the talk-aloud protocol and 1 minute and 8 seconds for the retrospective report which has been transcribed to 762 words and 159 words respectively.

The users A and B who were interviewed face to face took approximately similar times in using the DucView tool. User C finished using the tool earlier. User B's transcript size was the least among the 3 users.

User	Time taken for talk aloud protocol	Words in the transcript in he talk-aloud protocol	Time taken for retrospective reports	Words in the transcript in retrospective report
A	6 mins and 26 seconds	329	33 seconds	50
B	5 mins and 09 seconds	700	1 mins and 33 seconds	246
C	6 mins and 08 seconds	598	1 mins and 59 seconds	275

Table 2: Data about users using the GATE tool

From Table 2, for user A, the recording was 6 minutes and 26 seconds in the talk-aloud protocol and 33 seconds for the retrospective report which has been transcribed to 329 words and 50 words respectively.

For user B, the recording was 5 minutes and 9 seconds in the talk-aloud protocol and 1 minute and 33 seconds for the retrospective report which has been transcribed to 700 words and 246 words respectively.

For user C, the recording was 6 minutes and 08 seconds in the talk-aloud protocol and 1 minute and 59 seconds for the retrospective report which has been transcribed to 598 words and 275 words respectively.

User C's transcript size was the least among the 3 users.

4.2 Usability Metrics

The currently existing usability metrics as per ISO standards are dependent on the user [8]. The 10 usability metrics considered are efficiency, effectiveness, satisfaction, productivity, learnability, safety, trustfulness, accessibility, universality and usefulness [8]. The dependence of these usability factors on user is illustrated in one of the parameters called task effectiveness.

Task Effectiveness (T.A.) = (Quality * quantity)/100

Quality is the proportion of the goal achieved and hence is dependent on the number and set of users. It does not give a score which can be applied for any set of users. The definition of the goal can be subjective.

To overcome, these challenges, a set of 5 objective metrics have been proposed which are fair selection of users, cognitive capability of users, impact of users on other users, focus on the task by the users and motivation of the users. These 5 metrics provide a score that give an idea about the usability of the interfaces of the tools. As, they are calculated on the users, it is assumed that these 5 objective usability metrics can be applied on any set of tools. The metrics can be calculated over any set of users unlike previous usability metrics which are dependent on a set of users.

2 more usability metrics have been proposed which are bias of the experimenter and reliability of the dataset. They are subjective in nature.

4.2.1 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) is a 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.

ROUGE score is obtained by the comparison of a summary with a set of reference summaries [12]. Normally, it can be used to evaluate the summarization quality of automated systems [12].

Here ROUGE score has been used to see how the precision, recall and F measure for the summary of one user of one annotation tool is related to the summaries of other users for the same annotation tool.

Low scores of precision, recall and F measure and different scores for different users indicate that the summarization capability of the other users are different to the reference user. This indicates the long term memories and short term memories of the users are different as summaries can be considered to be a manifestation of the long term memory of the user. Thus, the users were selected fairly for using the annotation tool as they have different cognitive capabilities.

Higher values of precision, recall and F measure and similar scores for different users will indicate that the users have similar long term and short term memories, hence their selection was biased.

Recall value in ROUGE = (number of overlapping words in other user's and reference user's summary)/ (total number of words in the reference user summary) [12]

Precision value in ROUGE = (number of overlapping words in other user's and reference user's summary)/(total number of words in the reference user summary) [12]

The F measure is the harmonic mean of precision and recall [14] which means that

F measure = (2*Precision*Recall)/(Precision+Recall) [14]

ROUGE-n score gives an idea about the number of n-grams overlapping among the other user's and the reference user's summary [12]. N-grams refer to contiguous sequence of n items in a text [15].

ROUGE-L score provides an insight to the longest matching sequence of words in the other user's summary compared to the reference user's summary, using the technique of longest common subsequence (LCS) [12].

The ROUGE metric has been implemented in Python programming language using the rouge library[13]. The rouge library of version 0.3.1 has been installed in Python 2.7 in Fedora 20 Operating System. It provides the ROUGE-L, ROUGE-2 and ROUGE-1 scores [13].

The ROUGE score has been obtained for the users A,B and C in DucView and GATE tools respectively respectively.

4.2.1.1 ROUGE scores for user summaries in DucView Tool

The retrospective reports of the users A, B and C are considered as their summaries. Initially the

summary of user A is considered as the reference summary which is compared to the summaries of users C and B. The precision, recall and F measures have been obtained accordingly. Similar the summaries of B and C are used for reference.

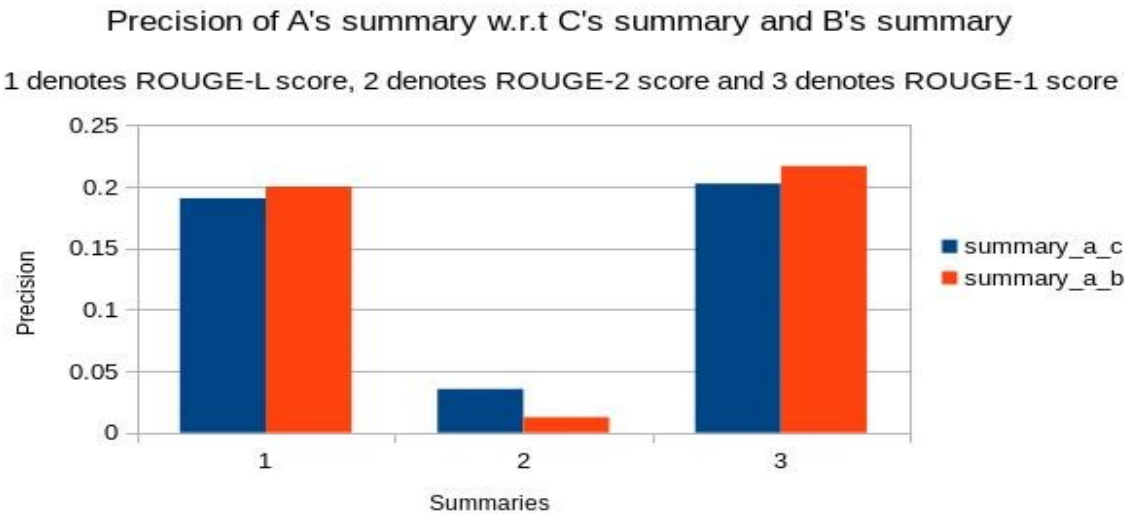

Fig. 1: Precision values of of user C's summary and B's summary with reference to A's summary
summary_a_c denotes C's summary with reference to A's summary
summary_a_b denotes B's summary with reference to A's summary

It indicates that the number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. It also gives an idea that A's summary size is moderate compared to C's and B's summaries due to the denominator of the precision score. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different.

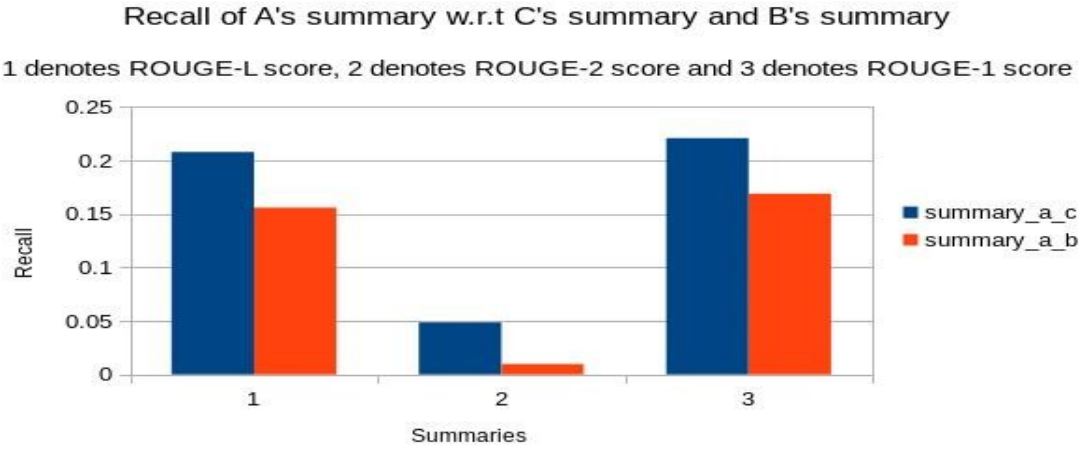

Fig. 2: Recall values of of user C's summary and B's summary with reference to A's summary
summary_a_c denotes C's summary with reference to A's summary
summary_a_b denotes B's summary with reference to A's summary

Fig. 2 indicates that the number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. Higher height of the blue column indicates that size of the C's summary is the highest as it is in the denominator of the recall score. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the recall values are different.

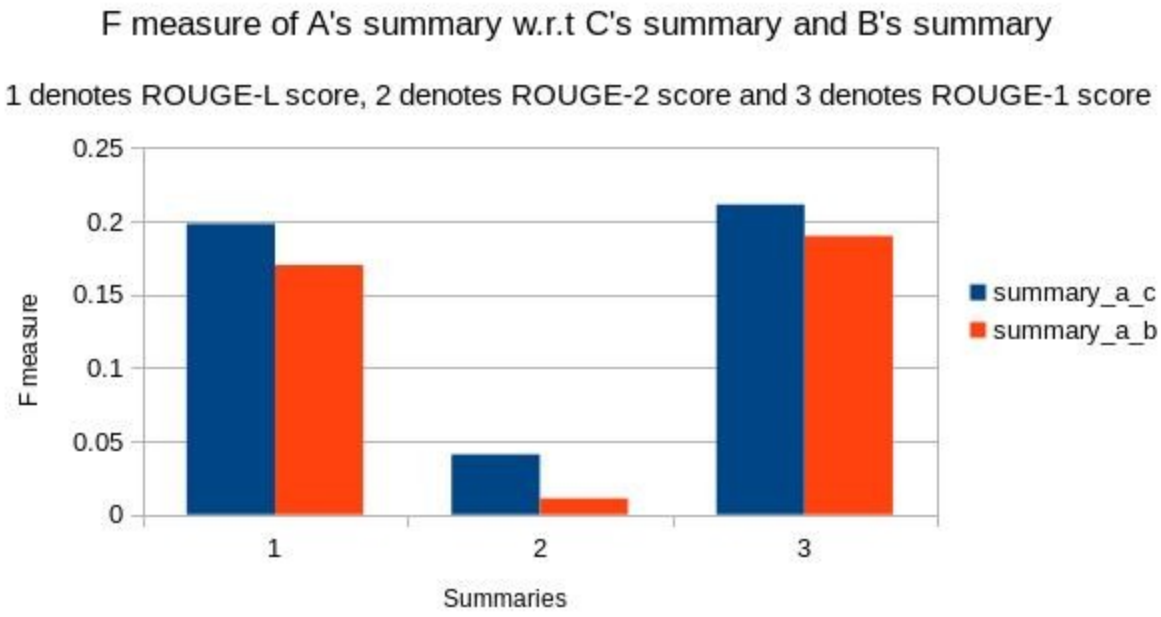

Fig. 3: F measure values of of user C's summary and B's summary with reference to A's summary
summary_a_c denotes C's summary with reference to A's summary
summary_a_b denotes B's summary with reference to A's summary

It gives an idea on the precision and robustness of the summarization. The number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. The F measure for the summary of user B with respect to A is higher than the recall value which indicates a comparatively robust summarization of user B. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are different.

From the precision, recall and F measure values of user A's summary compared to user C's and user B's summary, it can be concluded, that the summarization capability of A is different from B and C's summarization capability.

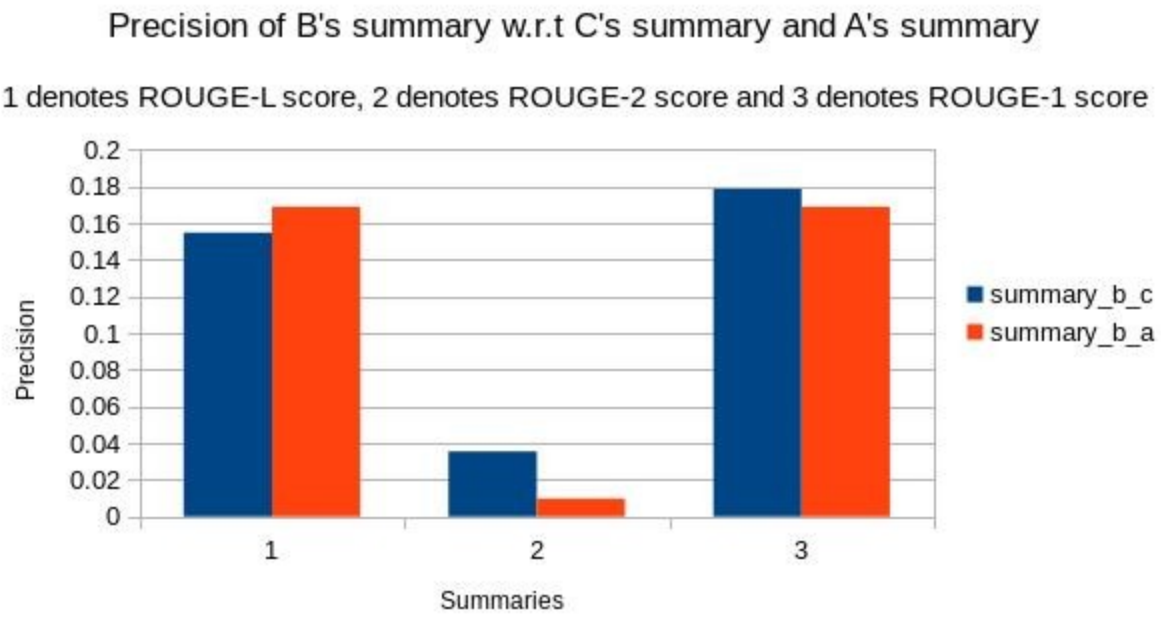

Fig. 4: Precision values of of user C's summary and A's summary with reference to B's summary
summary_b_c denotes C's summary with reference to B's summary
summary_b_a denotes A's summary with reference to B's summary

It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. It also gives an idea that B's summary size is least compared to C's and B's summaries due to the denominator of the precision score. Precision values are lower compared to the reference being A's and C's summaries. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different.

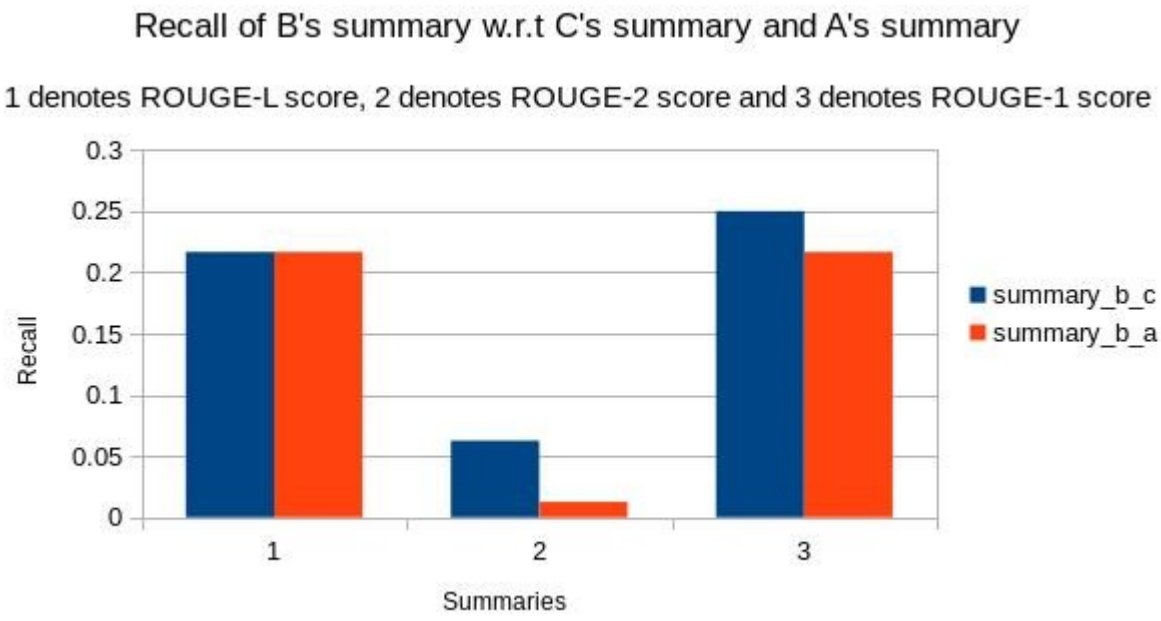

Fig. 5: Recall values of of user C's summary and A's summary with reference to B's summary
summary_b_c denotes C's summary with reference to B's summary
summary_b_a denotes A's summary with reference to B's summary

It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-2 and ROUGE-1, the recall values are different. ROUGE-L scores are similar in recall but since the scores are low, it could be coincidental that similar number of longest common subsequences are present in B's and C's summaries and B's and A's summaries. The number of unigrams overlapping in B's and C's summaries is more than that of B's and A's summaries.

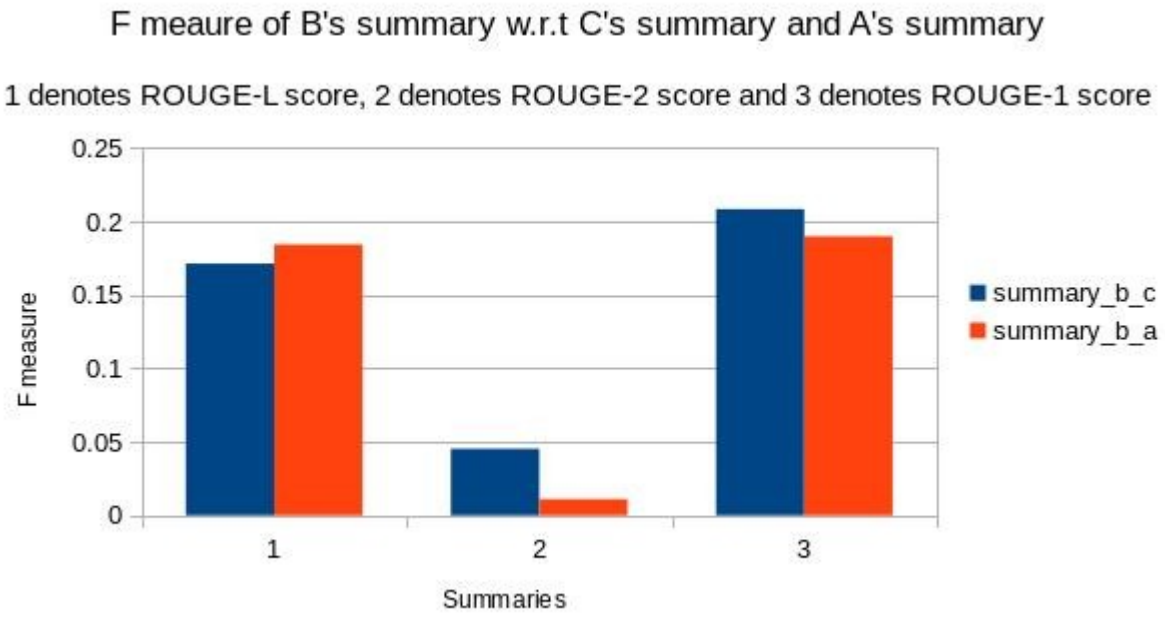

Fig. 6: F measure values of of user C's summary and A's summary with reference to B's summary
summary_b_c denotes C's summary with reference to B's summary
summary_b_a denotes A's summary with reference to B's summary

It means that the summarization of C and A is comparatively closer as more robust scores are being obtained in ROUGE-2 and ROUGE-1 scores. It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are different. The number of unigrams overlapping in B's and C's summaries is more than that of B's and A's summaries.

From the precision, recall and F measure values of user B's summary compared to user C's and user A''s summary, it can be concluded, that the summarization capability of B is different from C and A's summarization capability.

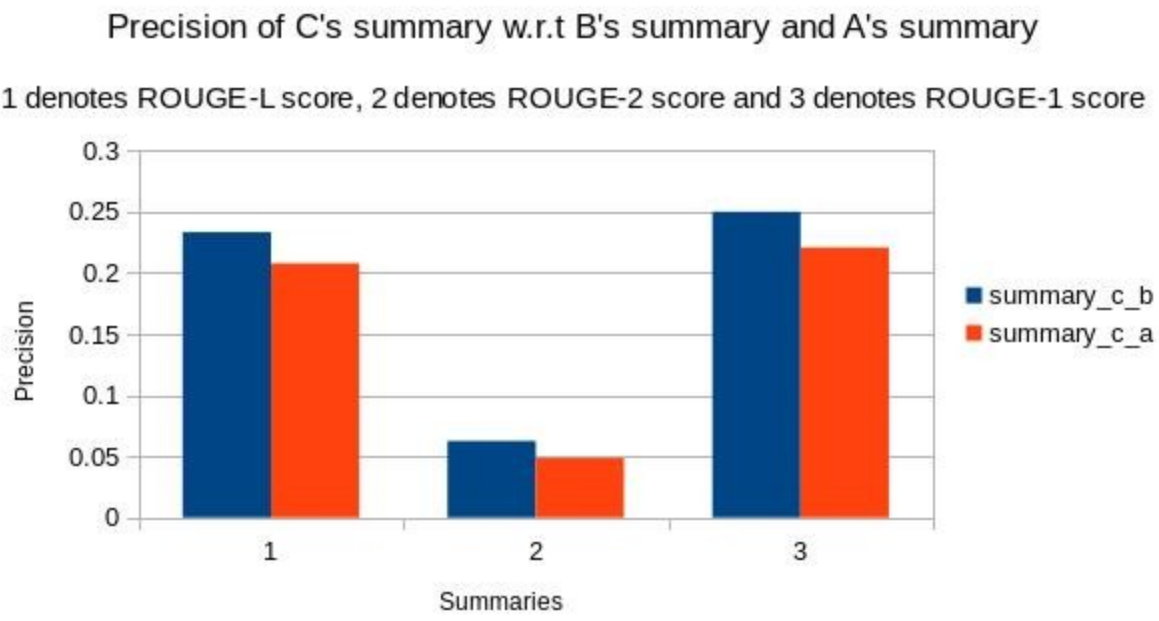

Fig. 7: Precision values of of user B's summary and A's summary with reference to C's summary
summary_c_b denotes B's summary with reference to C's summary
summary_c_a denotes A's summary with reference to C's summary

It indicates that the number of overlapping words in the user B's and A's summaries is much lower compared to C's summary. It also gives an idea that C's summary size is highest compared to B's and A's summaries due to the denominator of the precision score. Precision values are higher compared to the reference being A's and B's summaries. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different. The number of words overlapping between B's and C's is slightly more than that between A's and C's summary which gives a comparative idea of their summarization capabilities.

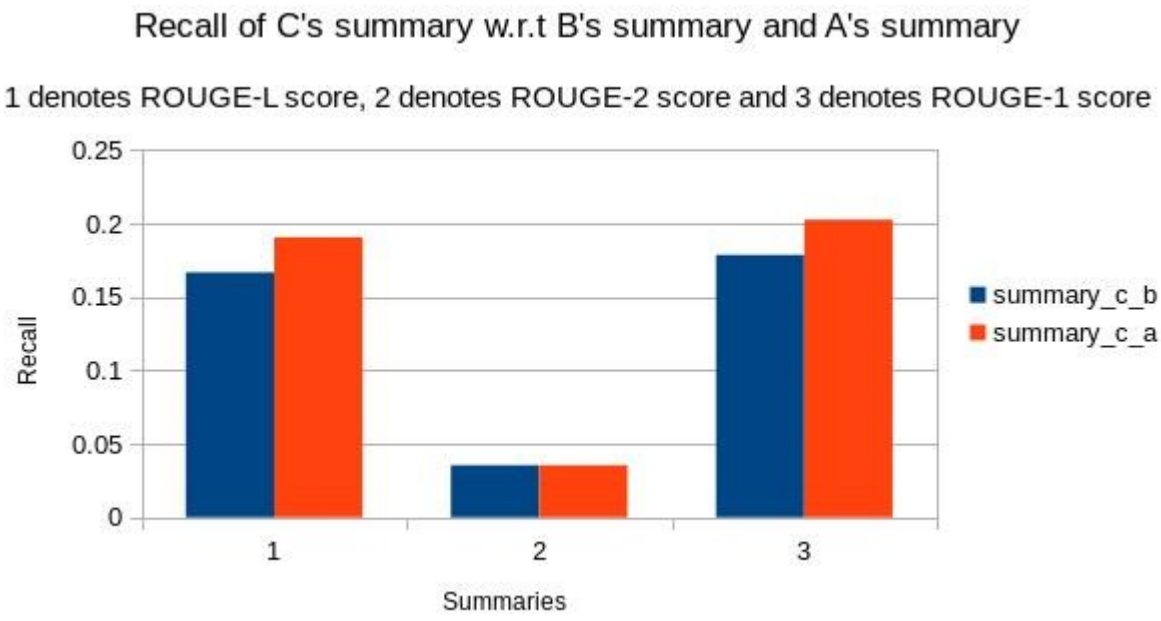

Fig. 8: Recall values of of user B's summary and A's summary with reference to C's summary
summary_c_b denotes B's summary with reference to C's summary
summary_c_a denotes A's summary with reference to C's summary

It indicates that the number of overlapping words in the user B's and A's summaries is much lower compared to C's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L and ROUGE-1, the recall values are different. ROUGE-2 scores are similar in recall but since the scores are very low, it could be coincidental that similar number of 2 contiguous words are present in B's and C's summaries and B's and A's summaries. The number of unigrams overlapping in B's and C's summaries is less than that of C's and A's summaries.

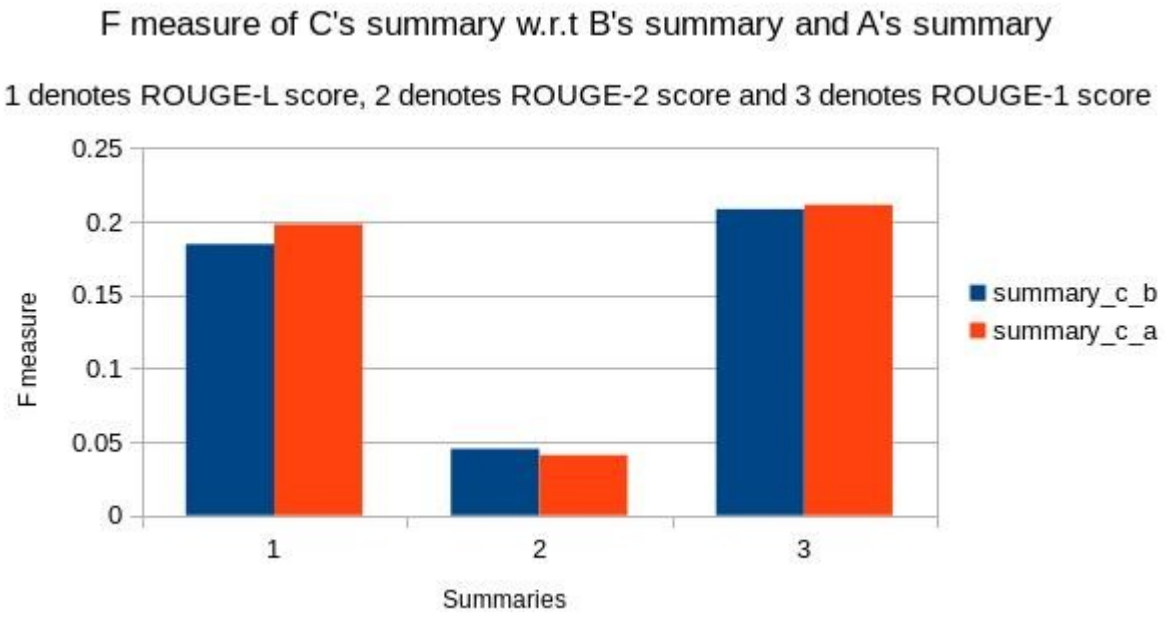

Fig. 9: F measure values of of user B's summary and A's summary with reference to C's summary
summary_c_b denotes B's summary with reference to C's summary
summary_c_a denotes A's summary with reference to C's summary

It means that the summarization of C and A is comparatively closer as more robust scores are being obtained in ROUGE-L and ROUGE-1 scores. It indicates that the number of overlapping words in the user B's and A's summaries is similar compared to C's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are slightly different.

From the precision, recall and F measure values of user C's summary compared to user B's and user A"s summary, it can be concluded, that the summarization capability of C is different from B and A's summarization capability.

This indicates that the long term and short term memories of A, B and C are different with respect to one another and their cognitive capabilities are different. Thus, it can be concluded that the selection of the users of DucView tool, A, B and C was fair without any bias from the objectively obtained ROUGE-L, ROUGE-2 and ROUGE-1 scores.

4.2.1.2 ROUGE scores for user summaries in GATE Tool

The retrospective reports of the users A, B and C are considered as their summaries while using the GATE tool. Initially the summary of user A is considered as the reference summary which is compared to the summaries of users C and B. The precision, recall and F measures have been obtained accordingly. Similar the summaries of B and C are used for reference.

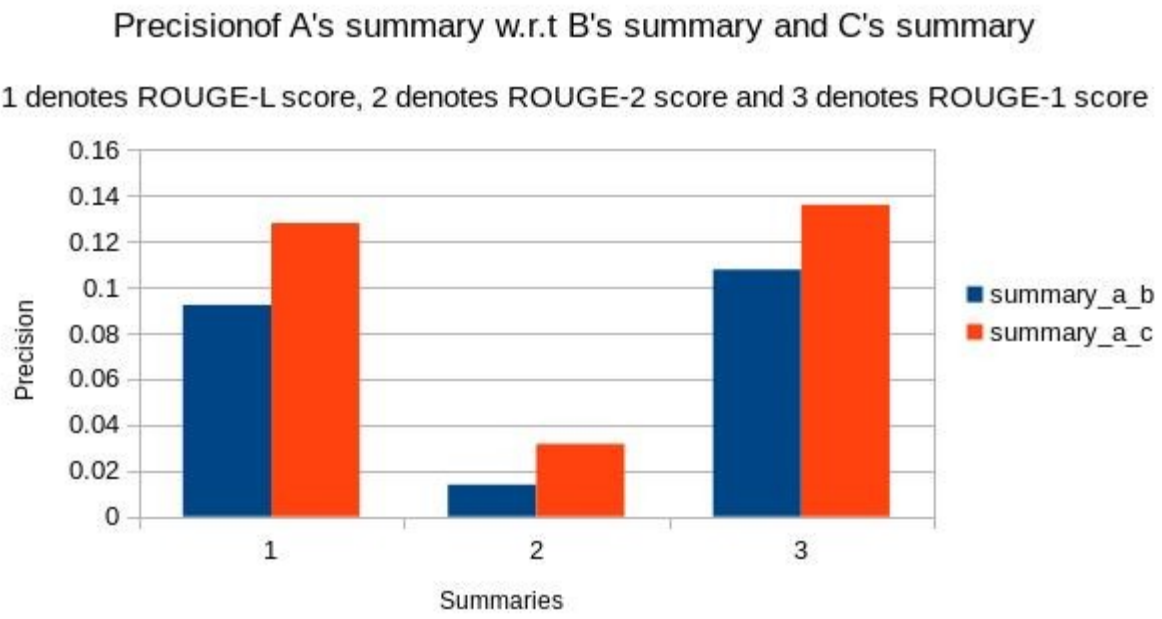

Fig. 10: Precision values of of user B's summary and C's summary with reference to A's summary

summary_a_b denotes B's summary with reference to A's summary

summary_a_c denotes C's summary with reference to A's summary

It indicates that the number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. The summaries are more distinct compared to the user's of DucView as the precision values are lesser than those of DucView users. It also gives an idea that A's summary size is least compared to C's and B's summaries due to the denominator of the precision score. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different.

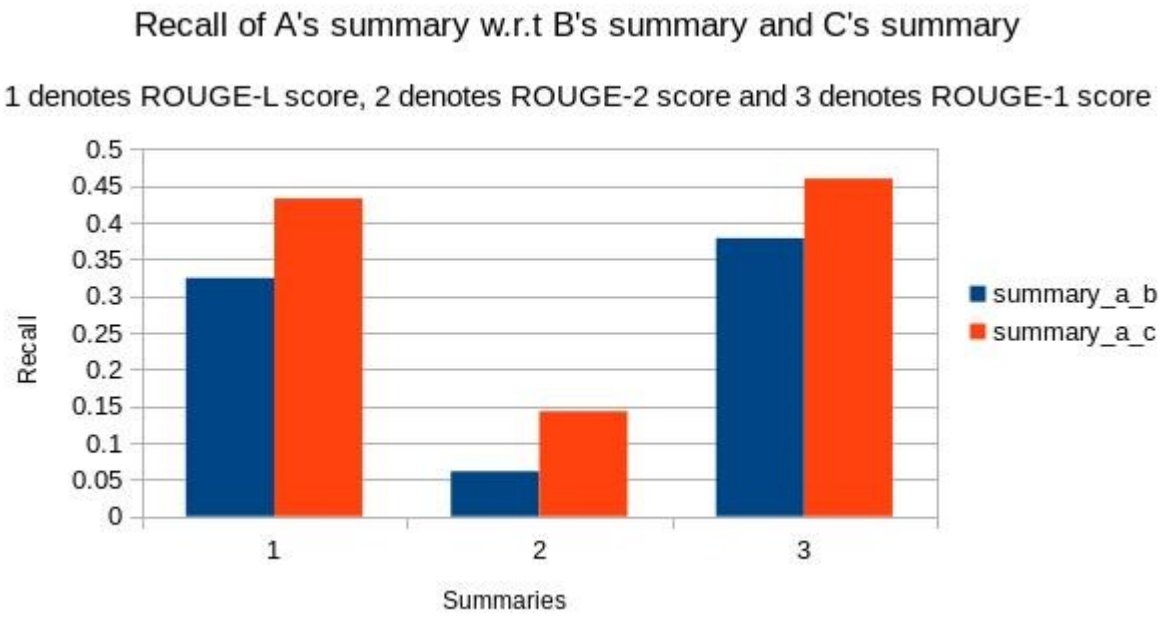

Fig. 11: Recall values of of user B's summary and C's summary with reference to A's summary

summary_a_b denotes B's summary with reference to A's summary

summary_a_c denotes C's summary with reference to A's summary

It indicates that the number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. The recall values are higher compared to similar recall values of DucView users. This could be because the users are taking comparatively similar strategies to annotate and also because the summary size of B and C are higher than that of A. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the recall values are different.

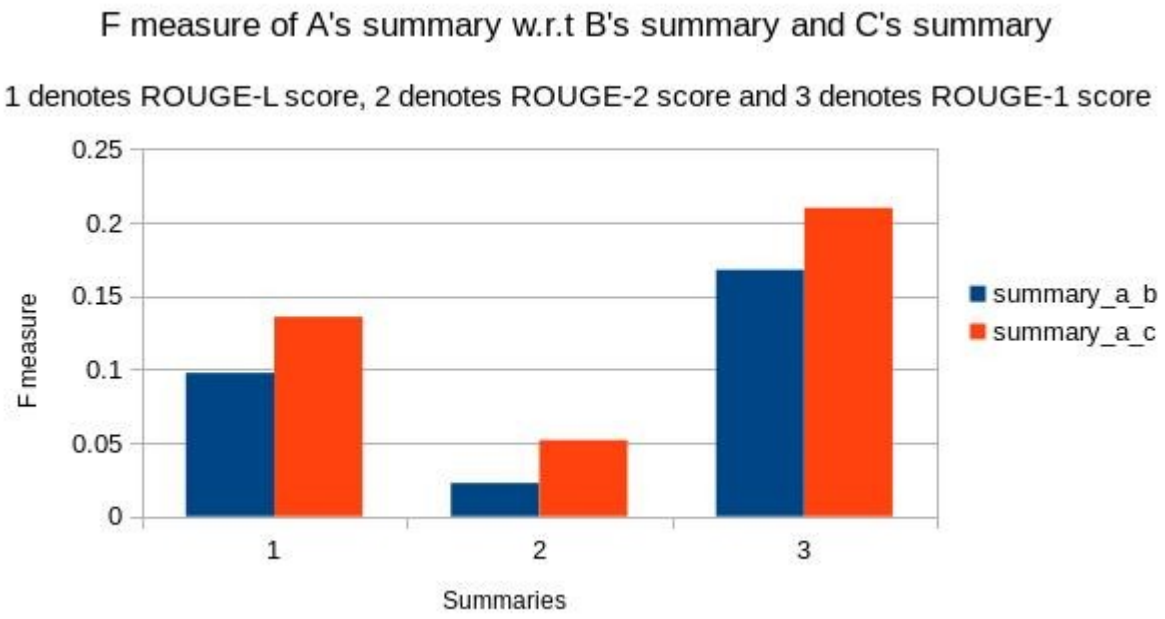

Fig. 12: F measure values of of user B's summary and C's summary with reference to A's summary

summary_a_b denotes B's summary with reference to A's summary

summary_a_c denotes C's summary with reference to A's summary

It provides an insight on the precision and robustness of the summarization. The number of overlapping words in the user C's and B's summaries is much lower compared to A's summary. The F measure for the summary of user B with respect to A is lower than the recall value which indicates a comparatively less robust summarization of user B. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are different.

From the precision, recall and F measure values of user A's summary compared to user B's and user C's summary, it can be concluded, that the summarization capability of A is different from B's and C's summarization capability.

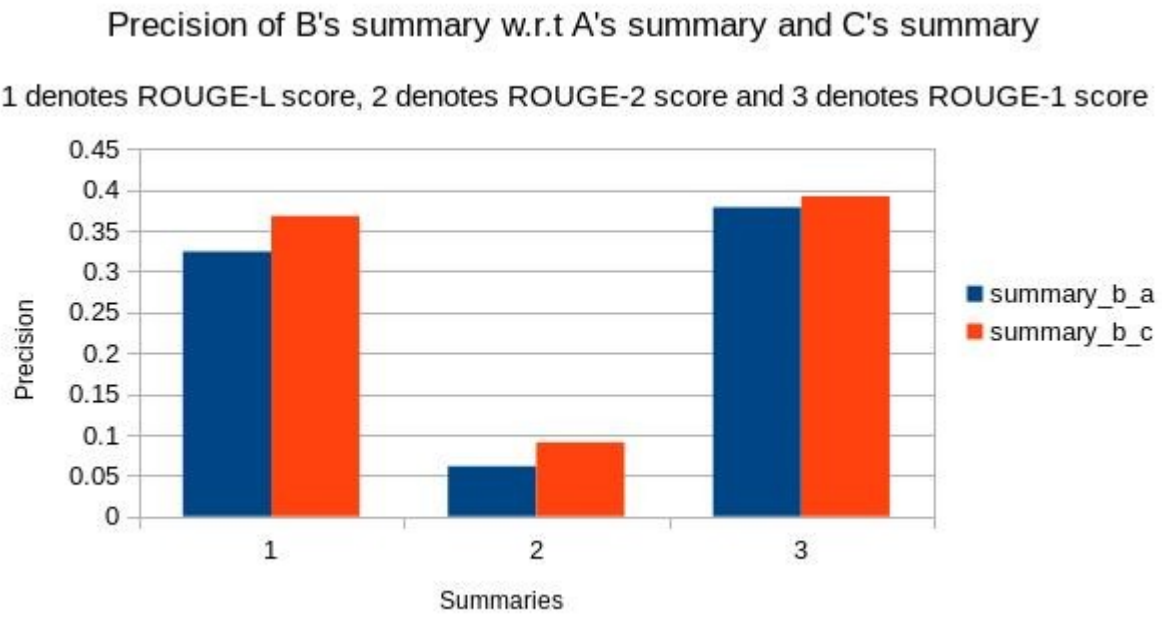

Fig. 13: Precision values of of user A's summary and C's summary with reference to B's summary

summary_b_a denotes A's summary with reference to B's summary

summary_b_c denotes C's summary with reference to B's summary

It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. The precision values of A's and C's summaries w.r.t B's summary are comparatively closer but overall the values are low which indicate dissimilar summarization. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different.

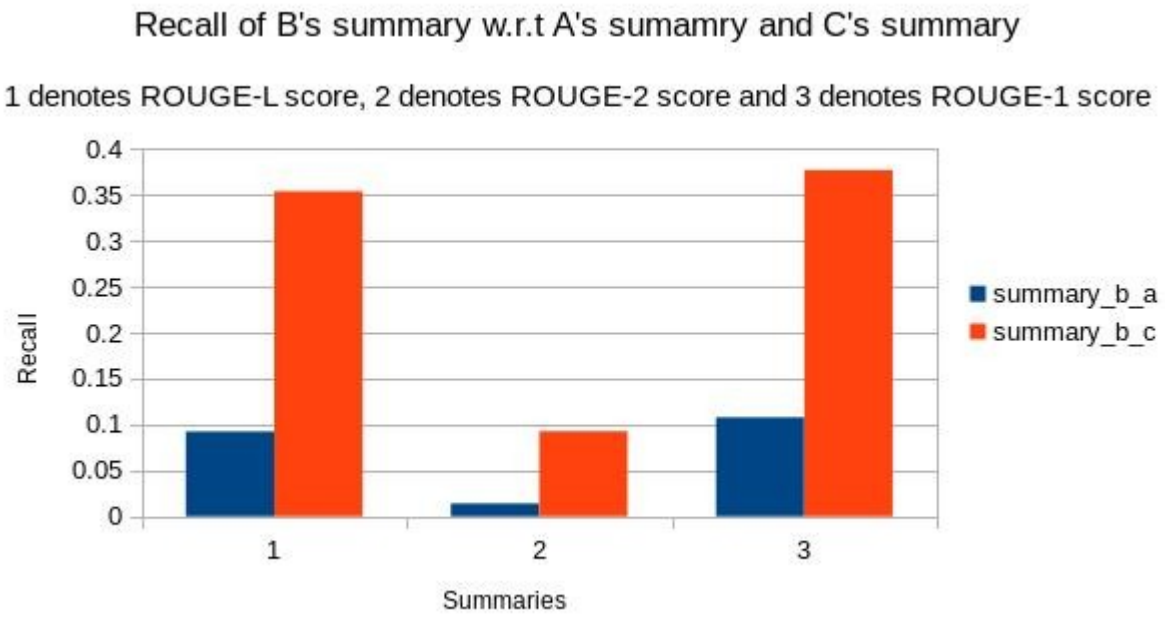

Fig. 14: Recall values of of user A's summary and C's summary with reference to B's summary

summary_b_a denotes A's summary with reference to B's summary

summary_b_c denotes C's summary with reference to B's summary

It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. The recall values of C's with respect to B's summary is much higher than tht of A's with respect to B's summary as C has generated a 5 times larger summary compared to A. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. The number of unigrams and bigrams overlapping in B's and C's summaries is more than that of B's and A's summaries.

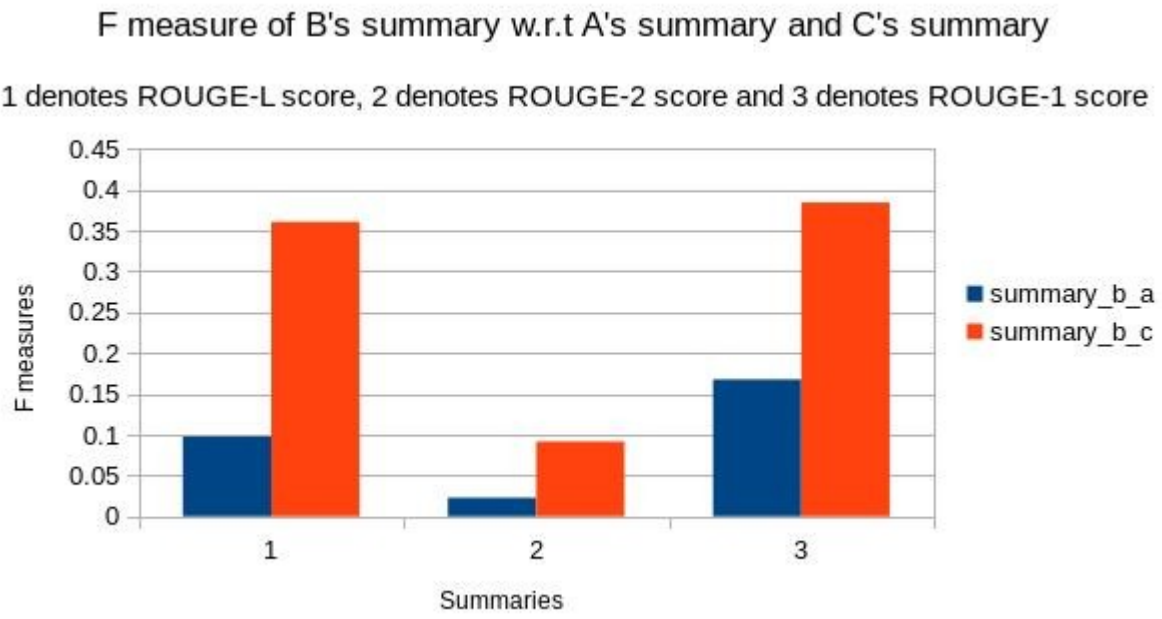

Fig. 15: F measure values of of user A's summary and C's summary with reference to B's summary

summary_b_a denotes A's summary with reference to B's summary

summary_b_c denotes C's summary with reference to B's summary

It means that the summarization of C and A are unique as much higher F measure scores are being obtained which can be due to the larger summary size of C with respect to A. It indicates that the number of overlapping words in the user C's and A's summaries is much lower compared to B's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are different. The number of unigrams overlapping in B's and C's summaries is more than that of B's and A's summaries.

From the precision, recall and F measure values of user B's summary compared to user A's and user C's summary, it can be concluded, that the summarization capability of B is different from A's and C's summarization capability.

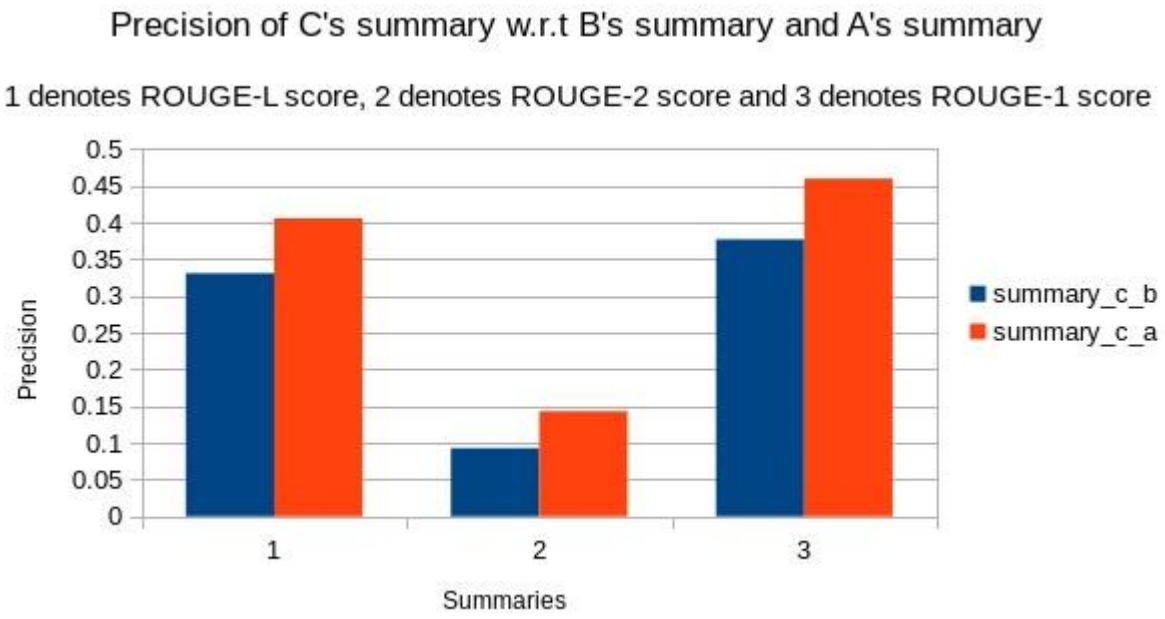

Fig. 16: Precision values of of user B's summary and A's summary with reference to C's summary

summary_c_b denotes B's summary with reference to C's summary

summary_c_a denotes A's summary with reference to C's summary

It indicates that the number of overlapping words in the user B's and A's summaries is much lower compared to C's summary. It also gives an idea that C's summary size is moderate compared to B's and A's summaries due to the denominator of the precision score. Precision values are much higher compared to the reference being A's and B's summaries. For ROUGE-L, ROUGE-2 and ROUGE-1, the precision values are different. The number of words overlapping between B's and C's is slightly less than that between A's and C's summary which gives a comparative idea of their summarization capabilities.

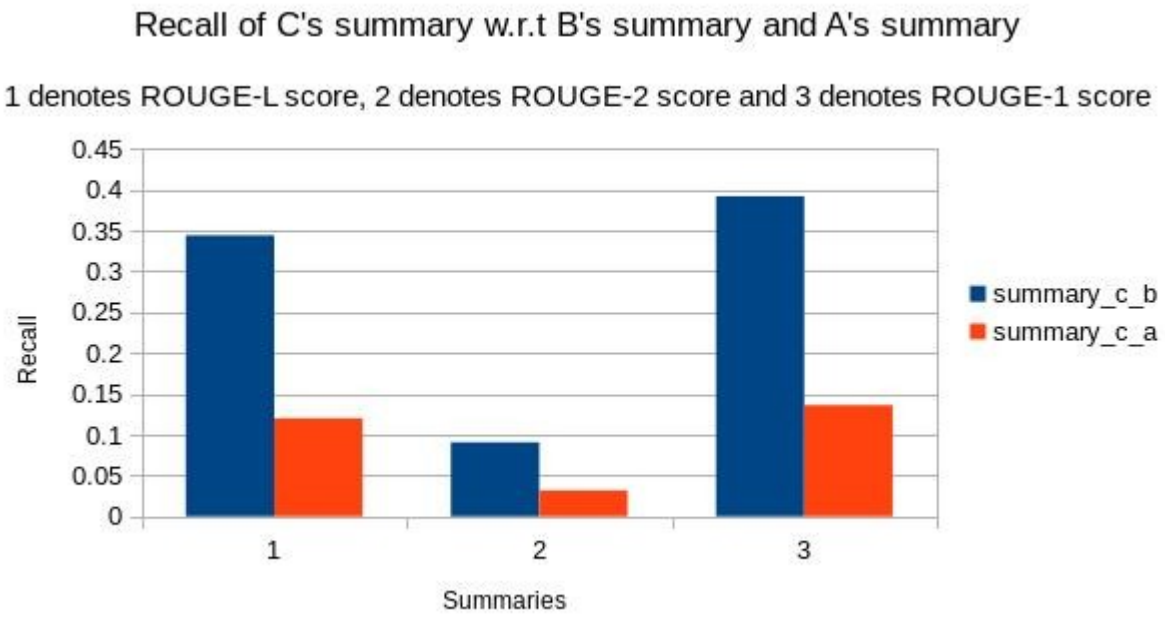

Fig. 17: Recall values of of user B's summary and A's summary with reference to C's summary

summary_c_b denotes B's summary with reference to C's summary

summary_c_a denotes A's summary with reference to C's summary

It indicates that the number of overlapping words in the user B's and A's summaries is much lower compared to C's summary. The f measures of B's summary with respect to C is much higher than that of A with respect to C. This indicates the dissimilar strategies of summarization and also that the summarization quality of A was the poorest among the 3 users. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the recall values are different. The number of unigrams and bigrams overlapping in B's and C's summaries is higher than that of C's and A's summaries.

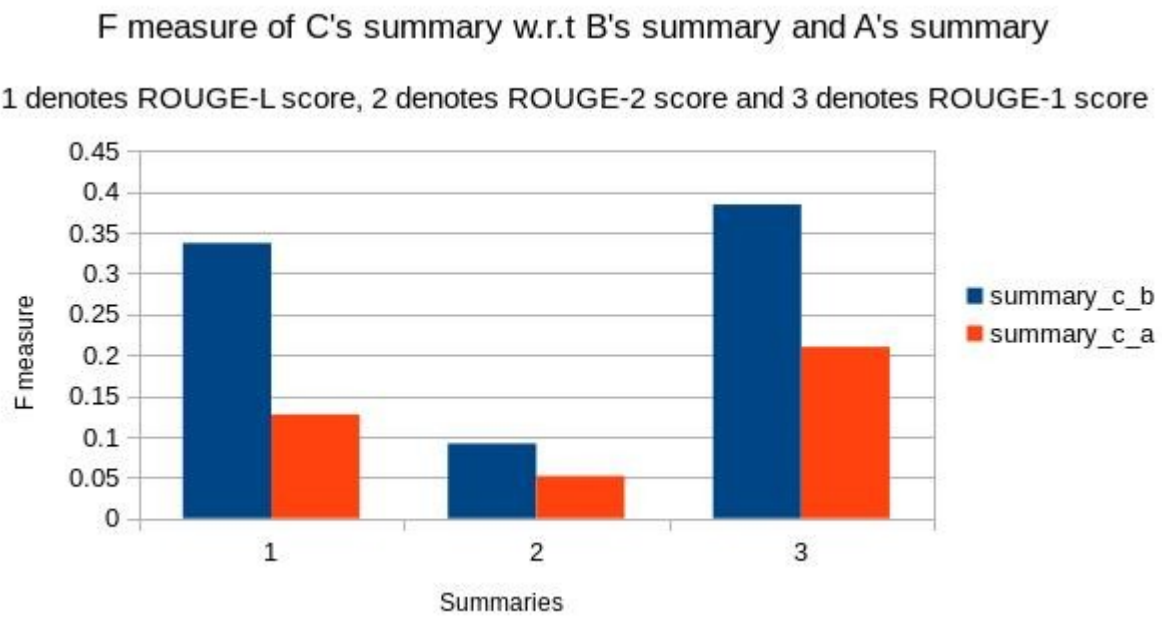

Fig. 18: F measure values of of user B's summary and A's summary with reference to C's summary

summary_c_b denotes B's summary with reference to C's summary

summary_c_a denotes A's summary with reference to C's summary

It means that the summarization of C and A is unique and different due to the wide difference in ROUGE-L, ROUGE-2 and ROUGE-1 scores. It indicates that the number of overlapping words in the user B's and A's summaries is different compared to C's summary. The longest common subsequence is closer to the uni-grams which indicates that it is not much impacted by 2 contiguous words in the summary due to lower score of the bigrams. For ROUGE-L, ROUGE-2 and ROUGE-1, the f measure values are slightly different.

From the precision, recall and F measure values of user C's summary compared to user B's and user A"s summary, it can be concluded, that the summarization capability of C is different from B's and A's summarization capability.

This indicates that the long term and short term memories of A, B and C are different with respect to one another and their cognitive capabilities are different. Thus, it can be concluded that the selection of the users of GATE tool, A, B and C was fair without any bias from the objectively obtained ROUGE-L, ROUGE-2 and ROUGE-1 scores.

4.2.2 Cognitive capability of users

The above obtained ROUGE scores from the retrospective reports provide an insight in the cognitive capability of the users. Another measure of cognitive capability of users can be based by the ratio of the number of words in talk-aloud protocol divided by the duration of the talk- aloud protocol.

For user C of DucView, the ratio is 762/5.6 = 136.071 units

For user B of DucView, the ratio is 485/7.15=67.832 units

The ratio gives an idea that C has better cognitive capability compared to B. User C is actually more experienced than user B and the transcript data is more useful for analysis. Thus the conclusion can be obtained from this score.

4.2.3 Impact of users on other users

A bag of words which are apologetic or dithering like ‘sorry, umm’ have been constructed. Users A and B were interviewed in front of one another. User A was interviewed first followed by user B. It has been observed that User A of DucView says sorry 9 times out of 457 words in talk aloud protocol while for user B, the frequency is only 3 times. It may be construed that user A was more impacted in sharing his thoughts in front of B and was nervous to make mistakes. User B learnt the mistakes of user A while listening to his interview and thus repeated fewer mistakes. Still User B was slightly impacted as his word frequency from the bag of apologetic words is present although at a much lesser extent than that of A.

The context of the words in the bag of apologetic and dithering words need to be analyzed for better understanding of the impact. The creation of the bag words is also another area of research as it needs to capture the apologetic and dithering words for a large set of user, who may express these qualities in different words. The inherent shyness of the users have not been considered in the impact of one user on the other. The measure depends on which user start the interview and which user ends the interview.

4.2.4 Focus on the task by the users

The focus on the task can be measured by a mechanism similar to Section 4.2.3. The weightage of the off-topic words can be calculated from the total number of words in the transcript. The bag of off-topic words need to be constructed from the transcripts that has the largest number of words as it provides more information than the other transcripts.

Focus on the task is a major challenge for the subject, as there are occasional emotional outbursts when the subjects face a problem during task execution which have not been investigated in the Ericsson and Simon paper[9]. A guided conclusion could have improved the focus which is not considered due to the talk-aloud protocol.

4.2.5 Motivation of users

The ratio of the number of words in talk-aloud protocol divided by the duration of the talk- aloud protocol can also be used also for measuring the motivation of the user to find which transcript should be given more importance.

4.2.6 Bias of the experimenter

It can be subjectively evaluated. Bias is introduced while removing noise while transcribing the audio recordings. The bias can be due to perception of noise or due to domain knowledge of the researcher. The researcher may remove some knowledge from the transcript thinking it is not relevant to the annotation while that data might have provided an insight in the usability study.

4.2.7 Reliability of the dataset

As user reviews from online websites have been considered, they might not be reliable as the identity of the users are unknown and often the usability ideas can be misleading as the same user can comment under multiple names in the websites and the researchers will have no idea of that as the websites abstract the identity of the users giving their reviews.

4.2.8 Weightage of the generic usability issues

A comparative analysis of the parameters mentioned in Sections 4.2.1 to 4.2.7 is necessary to understand how each of the generic factors influence the usability study of the users and whether an overall usability metric can be devised from each of the generic usability metrics.

4.3 Identification of Most Important Sub-task using Hierarchical Task Analysis

The data for task analysis was collected from the 6 users using the 2 NLP annotation tools GATE and DucView. The data was obtained by verbal protocol analysis using the talk-aloud protocol, data from online user review of annotation tools and by self-reflection.

Analysis of the data obtained has been used in application of HTA to conclude the task analysis on using NLP annotation tools. The transcriptions of both the users provide the groundwork for finding the tasks and sub-tasks.

Analysis of the transcriptions show that both the user open the DucView tool by moving the mouse to the Linux terminal icon, clicking it and typing a command in the Linux terminal and then enter is pressed in the keyboard to open the tool. Then once the tool is opened they find the file icon, move the mouse toward its it and click the icon. On clicking, the new button of the specific application is opened which is clicked again. This leads to a set of files from which the correct file for annotation is obtained by moving the mouse to search the file, clicking on the correct file, moving the mouse to the open button and clicking on the open button. The phrases in the text are tagged by moving the mouse for

selecting the phrases, then to move the mouse to the new annotation unit button, click it. By this way, a label is added. Multiple labels can be added like this by a loop. A label can be changed if the user wants to by clicking on the label, moving the mouse to the change label button, clicking it, moving the mouse to the text box, type the new name, moving the mouse to the OK button and clicking OK. If the user wants to stop, he has to move the mouse to file icon, click it, move the mouse to save as icon, click it, give the desired name of the file to be saved, move the mouse to the save as button and then click it to terminate.

Then it is saved in view of the observation of the users of GATE tool of continuous backup after annotation. Ctrl-S is pressed in the keyboard to saved the annotation. The decomposition of the task into sub-tasks can be pointed out in tabular form where the sub-points are the steps or the operators. The table has been represented in the next page.

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
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

Table 3: Task Analysis of using NLP Annotation tools by Hierarchical Task Analysis

There are seven sub-tasks and the strategy to complete the task is to execute sub-tasks from 1 to 3 linearly. 4,5 and 6 tasks can be executed in a loop as multiple labels can be added. For simplicity in representation, the iteration is not considered. Step 5 is optional to provide the actual name of the label e.g. any Part of Speech, but for good quality annotation, the annotator has to execute step 5 efficiently. Step 7 ends the task of annotation. The condition to move from one sub-task to the sub-task is on successful completion of the previous sub-task. The exit sub-task can be done when then user wishes to exit.

The sub-task with the largest number of steps is considered to be the most important sub-task. Step 5 has the largest number of steps which is 8. Step 5 is optional for the purpose of using the tool. However, the problem is that using should be linked to the domain knowledge of the tool which is natural language annotation here. So, renaming the labels is essential for annotation.

The second most important sub-task is exiting the annotation tool with 7 steps as the annotation needs to be saved appropriately.

The identification of the most important sub-task is essential so that researchers can effectively train new annotators on only these important sub-tasks. It saves time and money of annotators salary.

4.4 Representation of Task Analysis in Herbal Tool

The screenshots from the representation of the Hierarchical Task Analysis using Herbal tool are given below. The screenshots of only 1 out of the 7 sub-tasks with its steps in the problem space are attached due to paucity of pages for the report for the remaining 6 sub-tasks.

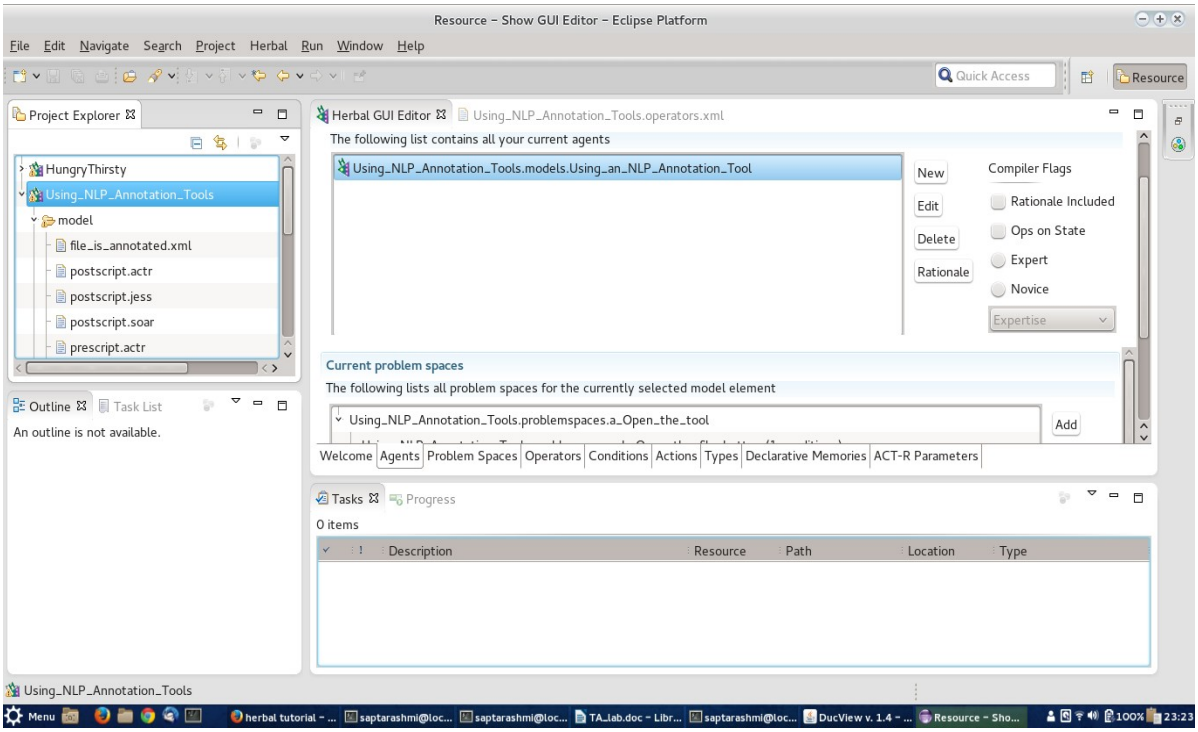

Figure 19: Representation of the agent and the problem space (continued in Figure 20)

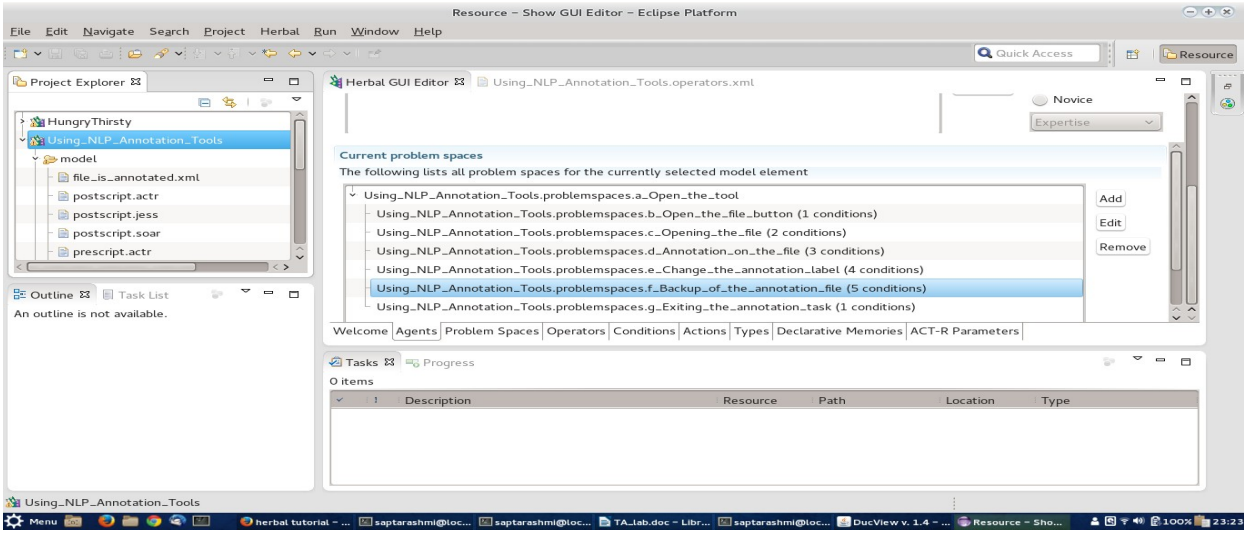

Figure 20: Representation of the agent and the problem space (continued from Figure 19)

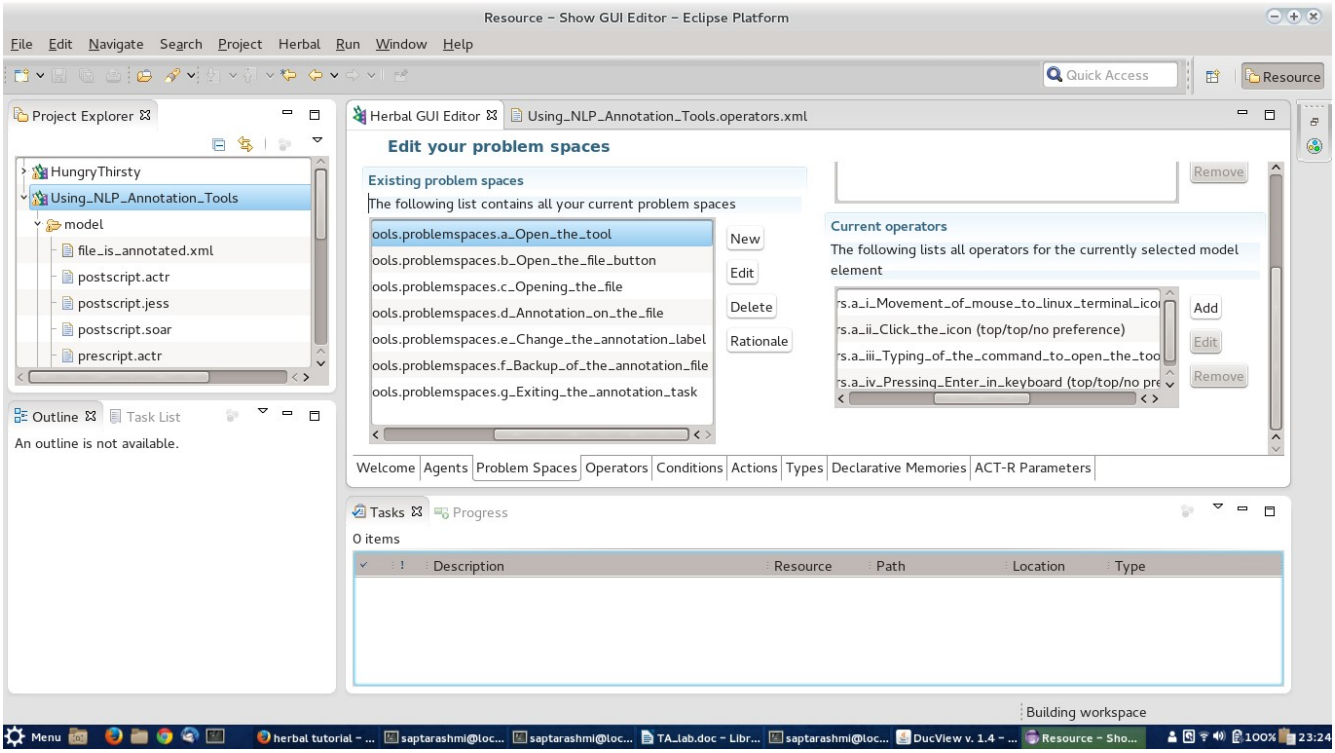

Figure 21: Representation of the sub-task 1 in the problem space and its operators

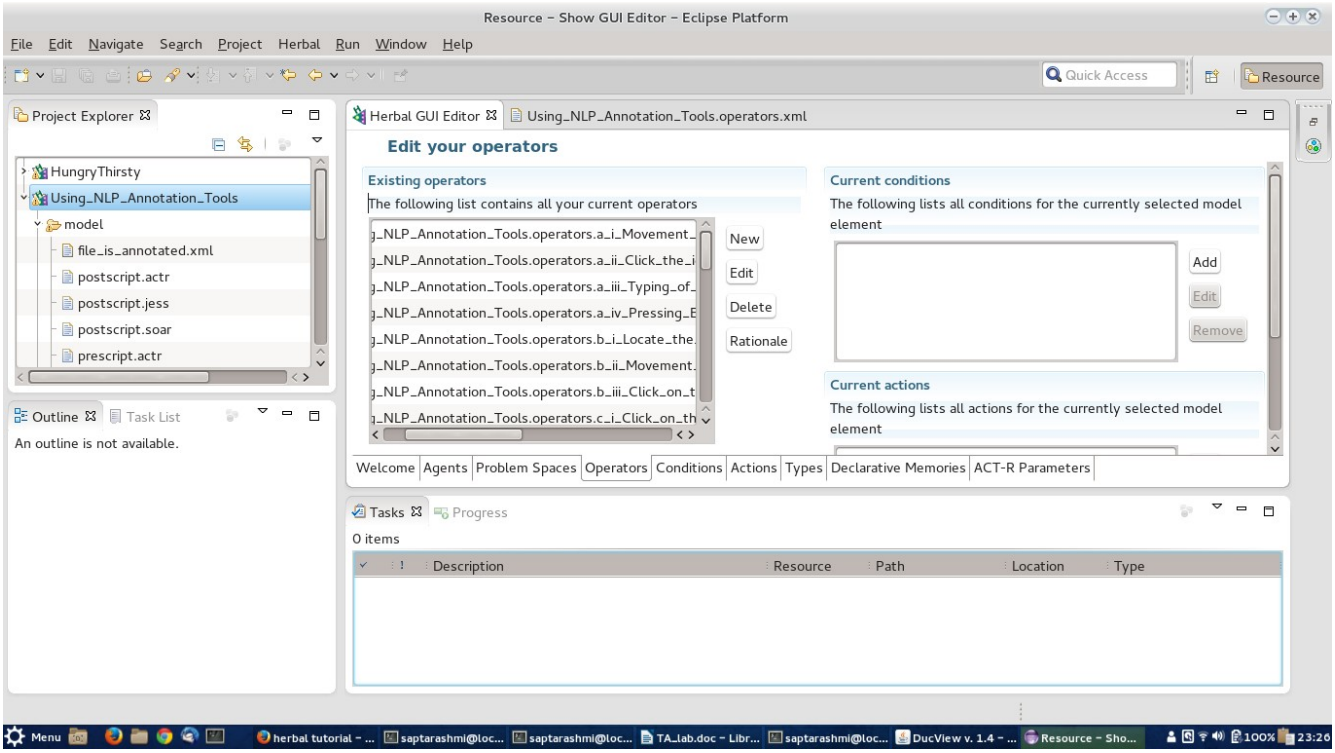

Figure 22: Representation of the operators in Herbal

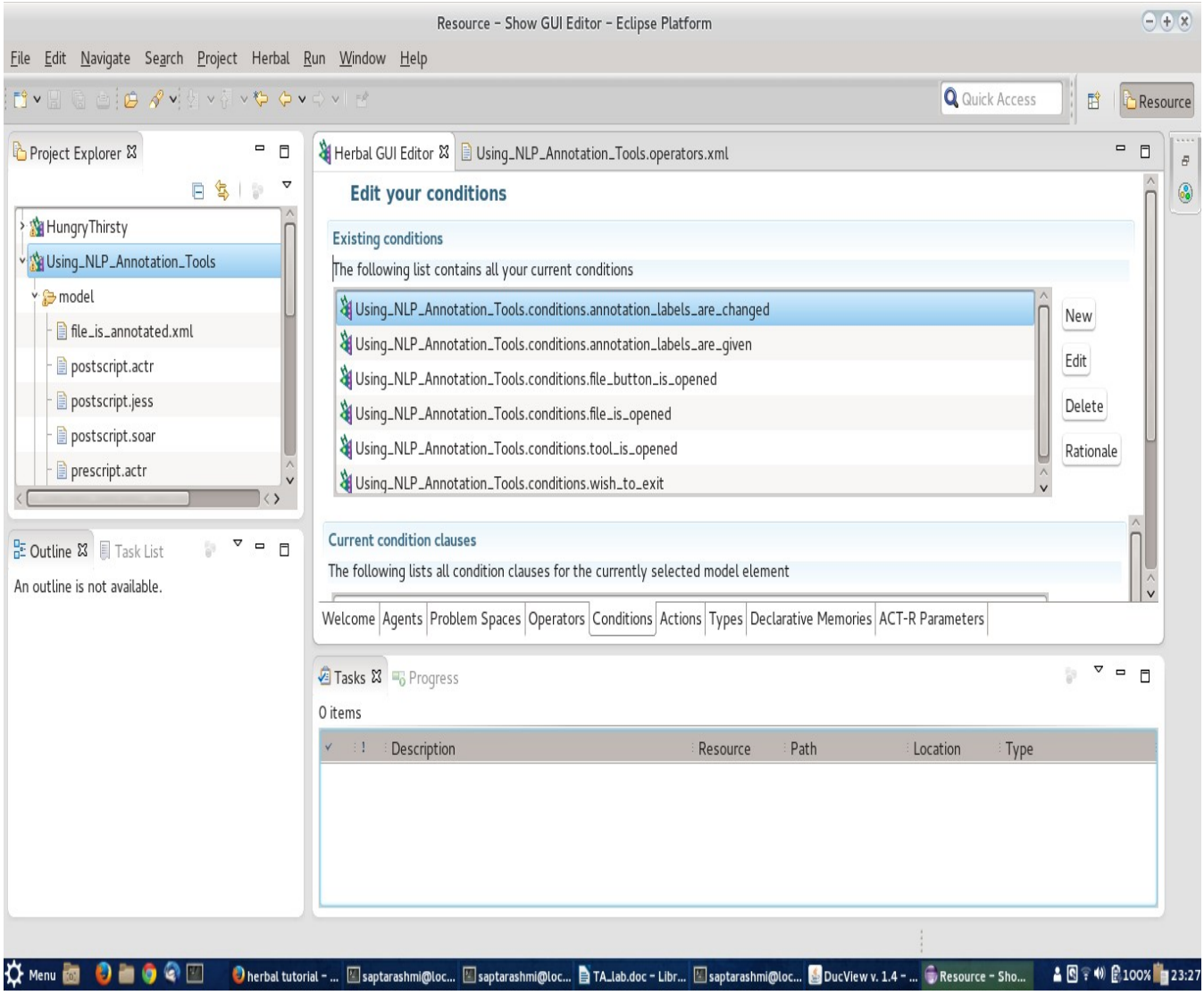

Figure 23: Representation of the conditions in Herbal

4.5 Usability Recommendation for the 2 NLP Annotation Tools

1) Controls of the software could be highlighted better with a better GUI(graphical user interface).

2) Selection of particular files is not highlighted clearly which indicate the need to provide footnotes with each function.

3) Confusion of the subjects illustrate more explicit instructions in the user manual. However, more subjects should have been tested to conclusively highlight the usability problems of the software.

4) 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.

5) The context in which the experiment is being conducted deserves deeper investigation as it controls the generation of data by the subjects if they cannot concentrate or other internal problems.

4.6 User Behavior

Some interesting psychological features have been observed among users like repeating different words confidently and repeatedly which the speaker does not mean to say actually and is unaware of.

One user (user A of GATE tool) mumbled while reading some portions of the text despite being prodded repeatedly to talk-aloud for those 14 seconds. It was interesting to note how the instructions were not followed by this user. Unsurprisingly from the ROUGE scores obtained in Section 4.2.1, it can be concluded that the user A's summarization was poor comparatively to other users B and C of GATE tool. While talking about the task, one user (user B of GATE tool) is taking actions and trying to reason the actions to own self.

4.7 User Strategies

The 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. Some users were self-explaining their answers in the talk-aloud protocol. Some users were reading the entire paragraph once by skimming and then starting to annotate.

Study of the strategies are required for better learning so that the tools can better service to users despite the strategies taken by them.

4.8 Identification of users who can potential drop-out from the task

The users of the GATE tool were initially given a quiz to understand their intent to participate. 32 quizzes were conducted. The users were given the option to attempt the quizzes two times. In the quizzes, the annotators applied co-specification annotation in example problems. A sample quiz is attached in Fig. 24. The data has been obtained for the class project with permission from the researchers in the ongoing Annotation project at NLP Lab, PSU.

A tutorial was provided to the to download the tool and understand its uses.

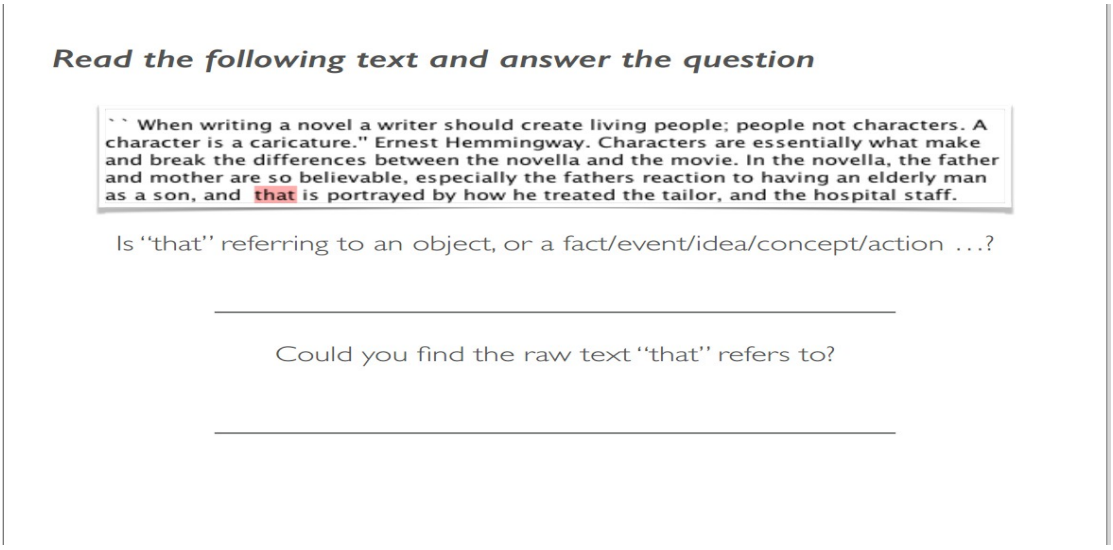

Fig. 24: Snapshot of the Quiz given to users of GATE tool

Users who stopped giving their feedback in the middle of the study (dropped out) have similar completion times of the warm-up task.

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.

If same quiz is used in future among a similar set of users, then potential drop-outs may be identified. This is because natural language annotation is not a fixed effect and can have different applications which require much broader analysis of different annotation tools. Otherwise the dataset becomes useless when user drops in the middle of a task. Such data can be used to act as dummy users.

However, this indicates, that the attempt is to bias the data positively in terms of annotators who are efficient rather than all types of annotators. An in-depth understanding of the bias is essential to see if it is harmful in the task of annotation. Also, the annotators who dropped out could be due to lack of motivation which can be evaluated from the usability metric of motivation of users in Section 4.2.7. Adequate strategies should be needed to keep the user motivated by periodic interactions with the users with interesting problems which can be considered as periodic warm-up activities. This will lessen the need of users to revise in future tasks which can lead to potential drop-outs.

4.9 Analysis of the result

i) The selection of 2 graduate students as subjects reflect their high short-term memory in the retrospective report which is consistent with the talk-aloud protocol. A general conclusion has been with a larger variety of subject selection.

ii) The retrospective reports ensure the completeness of the talk-aloud protocol, although the conclusion could be affected with more subjects, which brings a number of factors into play.

iii) The task objective is completed by both of the subjects in approximately the same time which hints a similar efficiency in their cognitive processing model.

iv) The users transcript sizes are different although interviews were conducted in the same time which highlights a difference in their mapping model of information to action despite showing similar efficiency in their task execution.

5. Discussion

Analysis of the result leads to the following points of observation:

i) It has been observed that the Hierarchical Task Analysis has many operators of mouse movements and mouse clicks. Thus the logical extension for cognitive processing would be to use keystroke level model (KLM) based on the data generated by these operators.

ii) The results may be considered general as one of the tools analyzed, the GATE tool is used for general text engineering while the DucView is specialized for pyramid annotation. This is because natural language annotation is not a fixed effect and there can be several types of language annotation like co-specification annotation and phrase annotation to name a few.

ii) Most of the operators have directionality and context attached to it. Like, a mouse can move to many buttons at different times or some buttons can be clicked at some circumstances only. This can be optimized with such operators being reused.

iii) The knowledge base is not reflected in this task analysis. Like, sub-task in step 5 is optional based

on the functionality of the software but for success of the annotation, the user has to exert step 5 to change the name of the labels by applying his or her knowledge of annotation. However, it requires the largest number of steps required to complete the sub-task, indicating that it is most important among all the sub-tasks.

iv) The roles of interfaces is not established in this task analysis although it is obtained from verbal protocol analysis or from the data of online user review. This is because HTA abstracts the interface.

v) Although the information, that the processing speed of the software is slow, can be obtained from the verbal protocol analysis, it is not reflected by HTA.

vi) The user interface of the Herbal tool could be improved as it does not allow reordering in the problem space and display the sub-tasks in the sorted order alphabetically. Also, it is not mentioned clearly that number cannot be given as input and any failure due to such reason can be corrected by changing the equivalent code that is being generated by the Herbal tool.

vii) The impact of the medium of interaction of the user interviews on the usability study needs to be understood. The two modes of communication with users have been face-to-face interactions as well as remote interviews on Skype.

viii) There are significant challenges on using the ROUGE score for the quality of summarization as a reference summary is needed. The quality of the reference summary is often not ascertained. A better metric must be devised to evaluate the quality of summarization.

6. Conclusion

It can be concluded that the usability of NLP annotation tools can be evaluated by a set of objective metrics namely
i) Fair selection of users: It can be ascertained by calculating the ROUGE score of the summaries of users by treating the retrospective reports as summaries to understand the difference in cognitive capability among users.

ii) Cognitive capability of users: Other than the ROUGE score, the ratio of the number of words in talk-aloud protocol divided by the duration of the talk- aloud protocol provides an insight the cognitive capability of users.

iii) Impact of users on other users: A bag of apologetic and dithering words have been created based on which the frequency of apologetic and dithering words have been calculated which gives an idea if users are impacted if they are interviewed together.

iv) Focus on the task by the users: A bag of off-the topic words can be created by studying the transcript with the largest number of words based on which the frequency of off-the-topic words have been calculated which gives an idea if users are getting defocussed.

v) Motivation of the users: The ratio of the number of words in talk-aloud protocol divided by the duration of the talk- aloud protocol can also be used to evaluate the motivation of the users.

There are 2 subjective usability measures identified which are:

i) Bias of the experimenter: It can be based on the domain knowledge of the transcriber, which is the researcher in most cases or due to noise in the audio which is being transcribed.

ii) Reliability of the dataset: User reviews in online websites can be unreliable as the data cannot be verified and the same user can give multiple reviews with multiple user names which cannot be identified.

A set of 5 usability recommendations to improve the interfaces have been proposed.

The most important sub-task in natural language annotation by these tools have been identified by HTA based on the number of steps. A detailed specification of 7 sub-tasks, 32 operators and 6 conditions has been obtained by HTA and mapped in the Herbal tool.

The users were not told any strategy in the talk-aloud protocol. The users took unique strategies while using the tool like reading the text line by line or skimming a paragraph and then starting to annotate.

Distinct user behavior was observed while studying the transcripts.

A proposal of identifying the users who can potentially drop-out while carrying the annotation, based on identifying a threshold completion time from all the completion times of the warm-up activities of the users of a tool have been suggested.
The verbal reports obtained from the transcription can be considered as data. The accounting of the verbal recordings and detailed explanation to the generation of reports is similar to other data processing. Retrospective reports ensure the completeness of the talk-aloud protocol. A deeper understanding can be obtained of the cognitive processes based on the instructions of the protocol to talk-aloud all the processes during the task execution. The conclusions in the experiment are consistent with the Ericsson and Simon paper[9].

7. Future Work

Keystroke level model can be used to measure the task timing. User data should also be collected by eye tracker and key stroke tools like RUI. Challenges of user availability and participation needs to be addressed.

More NLP annotation tools and platforms having different functionalities need to be reviewed for their usability features. Brat is a web-based annotation tool to capture annotation of social media data. No current usability resource is available for Brat. Context of annotation can be uncertain in Brat which has been used for many kinds of annotation, including transactivity detection as shared by Prof. Carolyn Rose, Professor at the School of Computer Science, Language Technologies Institute and the

Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, U.S.A. while giving a talk at the NLP Colloquium at Pennsylvania State University in 2018.

Usability review should be done on Amazon Mturk too. It is a platform for crowd sourcing where Human Intelligence Tasks (HITs) are given to workers. Human behavioral experiments are challenging to be conducted on Mturk as users can drop out in the middle of the main task, which leads to a waste of time, money and data. Users can drop out due to lack of appropriate motivation. Also, data from turkers are often not reliable as the turkers are not motivated enough with their only objective being obtaining the reward for the task in Mturk.

Automated transcription can be done using speech to text tools like CMU Sphinx. But the transcripts need to be reviewed manually as there can be errors in the automated transcriptions.. A think-aloud protocol can be used to study the impact of emotions of user's behavior on the usability study. However, several parameters like emotions and context have not been considered in the Ericsson and Simon paper [9] which provides the ground for future research work. Selection of subjects, reflecting a variety of parameters, will help in reducing bias in the result. The different user strategies can be adapted by the user interface to improve the user experience. The number of users need to be increased to obtain a generalizable result.

IRB approval is needed for publications and funded projects in this work as human users are interviewed for the purpose of the class project.

8. 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/

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

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

12. https://rxnlp.com/how-rouge-works-for-evaluation-of-summarization-tasks/#.XA_OYaBOnVN

13. https://pypi.org/project/rouge/

14. https://en.wikipedia.org/wiki/F1_score

15. https://en.wikipedia.org/wiki/N-gram

16. Appendix of the Book “Protocol Analysis” , by K. Andersen Ericsson and Herbert A. Simon, MIT Press, Cambridge, Massachusetts, 1993, http://acs.ist.psu.edu/ist521/papers/ericssonS93A.pdf
17. http://acs.ist.psu.edu/projects/herbal/herbal-tutorial.pdf

18. https://stackoverflow.com/questions/25491886/human-annotation-tool-for-corpora-in-nlp

19. https://www.quora.com/NLP-what-are-the-best-tools-for-text-annotation

20. https://www.researchgate.net/post/Does_Gate_allows_automatic_annotation_of_documents_using_domain_ontology

21. https://www.ibm.com/developerworks/community/blogs/e8206aad-10e2-4c49-b00c-fee572815374/entry/icm_to_do_lists_and_task_properties?lang=en

22. http://www.psy.gla.ac.uk/~steve/HCI/cscln/trail1/Lecture8.html

9. Appendix

It is to be noted that the transcriptions of the six users are provided here in a first person account. The transcriptions have been written manually from hearing the recording of the verbal talk-aloud protocol

for six minutes and retrospective report for two minutes approximately. The recording was stored in the sound recorder application of my cell phone which was listened to during the trancription.

Transcripts of Users of DucView tool

The transcripts of the 3 users A, B and C of the Ducview tool have been documented for analysis in the above experiments.

Transcription of User A during the concurrent verbal talk-aloud protocol for approximately 6 minutes

All right, now we have to do the Java thing again as in the practice. Do I get to pick any file from the computer? Oh sorry it is written, I could. That's nice. So, there is an initial text file and then I can mess with that file. It could be anything like a text or a pdf. Oh, it's working. But it's so slow. Here we go. Sorry, I do not know what that says in the prompt. Why is the prompt, hidden in a corner? It should have been displayed in the middle to draw attention of any user like me. What did I open? Sorry, I clicked on load. Ohhh!! Does new and load do the same thing? I see. There is nothing to load. I should have done new. Interesting, very interesting. This is going to be different. I thought it will come up as a message like invalid 2*2 UTF byte sequence. I am not quite a 100% sure on what's going on? Is that why, it appears to be slow, when the progress of the work is shown in a tiny corner rather than a big prompt? I feel like it should be a pop-up. Oh god, which file to take? One of this should do it. Oh wait, sorry, this is the new thing. Which one should be selected? Umm, I'm not used to such a touch-pad. Wish you could pick one file for me. I have to pick a file that works and also has readable text in it. Wait, I can go back to the folder during warm-up. Here we go. Okay, now I have to find some labels. Oops, I need to select some text. Sorry, should I select lot of text or no text? Let's go for lot of text. Oh, awesome. Now what can be the label? Umm. Let's go for cool label. That worked! It has text in it, and I can expand it and close it. Now, I need to add some contributors. Sorry, No, again!! I have to select text again. It's sort of slow. Add contributor. Okay, let me add it to the label. And, it's done. Anything else to be done?? Oh, it's still doing my add contributor thing. We have to get this to work. It's definitely adding it. It says it's adding it. It needs like a load bar, just in case of such scenario where I don't know if it is doing something or it's not doing something. It's good, otherwise. Can I select the comment? I hope so. No? Oh yeah sorry, I can. Now somethings are selected and they are like sub-labelled. Yeah, this thing surely needs a loading bar. All set and done.

Transcription of User A during the retrospective report for approximately 2 minutes

All right, Let's see. Getting into it was barely simple but slow. The new SCU file worked out well. It is a bit weird, that it is mentioned only a text file as I could pick a pdf or something else. It should by default show the all files. Adding a new label thing was pretty easy. It's kind of strange that the load a file button has no prompt showing the progress of the work, but is working silently in one corner of the screen. I would like if it had a loading bar or something, so that I could see whether it's kind of stuck.

Transcription of User B during the concurrent verbal talk-aloud protocol for approximately 6 minutes

Okay, I'm going to go here to the terminal. Up, up. Here is the command. Okay, the DucView folder comes here. And this one. Now I have the tool, opened. So, I have to load. Oh no, where should I go? I don't know! Let me check the manual. Umm, okay, I need to go for the new button, not load. So, now starts the document hunt. There's nothing here. Right, right. Damn it, where are the files? I need to get anything. Oh, there it is, pyramid files. It should be all files. Never mind. Now up, for all files. Which file to select? Umm. Umm. Okay how about this doc file? Awesome. Wait, it shows, invalid byte format, let's go again. File, pyramid, new. Ohh! I should find a compatible file for the software. True, true. So file, pyramid, new. I remember where to go. All right, file selected. Creating, creating. All right, got it. I think, increase the window size, here, here. So, this button now! Okay, select some text. Let's create some labels now. A bigger button would be helpful. This keyboard is difficult to use. Now, let's try to change some label. Hmm. It highlighted. Collapse? What did I do? Let's try to write in school in here. All right, so I named the phrase as school. Oh no, I need to select the thing. Where did it go? New SCU unit. Oh no, let's change the label. Let's see what I created. Okay, so two labels. All right, it is getting highlighted now. So, now I have labels, tags whatever. Once, I click them, they get highlighted. Interesting, interesting. So, now comes my comment. Where should I give the comment? Okay, here. Let's say school. That's nice. But wait, where are the prompts? 50 years of advancement in computer science and this is what I get? A tiny display on that side, it should be bigger. Otherwise, it's done. Okay, I've completed my task.

Transcription of User B during the retrospective report for approximately 2 minutes

So, I tried to find the file which has to be run in that software, in order to tag, highlight whatever I should do to understand the SCUs. Then I created tags, which could expand, but the button size could have been bigger and better. I can comment and play around with the phrases. I commented based on what I thought could be the main topic of the line. The software seemed to work somewhat okay. But it did work finally.

Transcription of User C during the concurrent verbal talk-aloud protocol for approximately 6 minutes

I've downloaded the DUCView jar file. I'm going to open that and it is opening. let's see where it is. Umm, open. Okay, it's open now. Going to file, pyramid and click on it to upload the text file which is the main step to view the text file. It's open and now I have to create structural units from this text file. Uhh, So let's see. It says 'Wales is following Scotland, and moving towards a call for an elected assembly with devolved powers'. Umm, so I guess, "Wales is following Scotland' could be a structural

unit in itself. Umm, 'and moving towards a call for an elected assembly with devolved powers', umm, I guess ' moving towards a call for an elected assembly; could be another structural unit and 'with devolved powers, as advocated by the Labour Party'. So 'advocated by the Labour party' could be one and I could label it as, umm, I could change the label of this structural unit to umm 'call was advocated by the Labour party', prss ok and move on. The next sentence is 'Labour has committed to the creation of a Welsh assembly, and party leader Tony Blair set out proposals for devolution, setting off a constitutional battle with the Tories'. Okay, so 'Labour has committed to the creation of a Welsh assembly', so this could be another structural unit. Umm, looks familiar to ' moving towards a call for an elected assembly' but for now, I will create as a separate one. Umm,' party leader Tony Blair' could be another structural unit and I could change this label to 'Tony Blair is the party leader;. Okay, moving on.' set out proposals for devolution' umm, could be, Tony Blair, so, I'll just say, 'he set up' perhaps, I'll just name him Tony Blair 'proposals'. Okay,' setting off a constitutional battle with the Tories'. Umm, I'll change this label to 'This set out a constitutional battle with the Tories'. Moving on, 'Conservatives oppose any form of devolution'. This could be another unit. 'and want to maintain a strong Welsh Office with a cabinet minister, believing that would produce the best results for Wales.' So 'and want to maintain a strong Welsh Office with a cabinet minister' could be one unit. And, another unit would be ' believing that would produce the best results for Wales.' and I could change this label to 'They believe, that would produce the best result for Wales'. Moving on ' Prime Minister John Major'. So, I could create this as an unit and change this label to 'John Major is the Prime Minister'. Umm, I'm typing this right now, pressing ok. 'and the Tories are against the establishment of a Welsh parliament'. ' against the establishment of a Welsh parliament', this could actually, umm, be a part of, the contributor to that the 'Welsh are opposing a form of devolution', add that as a contributor. ' Tories are against the establishment of a Welsh parliament which has eroded the usual support conservative legislators had received in Wales. 'Plaid Cymru, the Welsh nationalist party', so I"ll add that to this another one and change this label to ' Plaid Cymru is the Welsh nationalist party' . 'stepped up its campaign for equal rights to Welsh self-determination', this could be another, umm , another structural unit and I'll just change 'Plaid Cymru stepped up its campaign'. 'demanding equal constitutional treatment with Northern Ireland, and Scotland', this would be another unit. ' The British government is pressing ahead with plans to reform the structure of local government in Wales.' Okay, umm. This could be another unit.' It will establish an elected Welsh assembly, with law-making and financial powers', this could be another unit. ' to replace the current two-tier system of county and district councils with single-purpose, unitary authorities.' So, 'to replace the current two-tier system of county' till 'unitary authorities' could be another 'unit'. Okay. ' The government intends to set up 21 new authorities to replace the eight counties and 37 districts in Wales.' I guess, this could be another structural unit. 'Shadow elections to the new uginary u new unitary authorities will be held as early as next year.' Okay, I'll create this as a new unit. ' Implementation of the local government reform will take place in April 1995.' Just been said. Yeah, I think I have created enough SCUs. I'll save it. Uhh, I'll go to File and go to Save As and I'll save it as WalesDUCView.pyr file on my Desktop. And I've saved it.

Transcription of User C during the retrospective report for approximately 2 minutes

Okay, so I downloaded the DUCView jar file as well as the Wales DUCView text file and saved them on my desktop. And then, I opened the jar file. Then I went to File and I opened, uhh, I went to pyramid, uhh, the text file. Umm, and after it was loaded, I looked at the first few lines to kind of get an idea what, of the paragraph is about, and think of how I can divide this into structural units. Uhh, so I,

my approach was to read each sentence one by one and try to think of dividing that into uhh, different parts, that provide some kind of information, some kind of phrases. That's what I did. I did one by one, read each sentence and tried to divide them into phrases that would independently provide some information to whole paragraph. So I did that one by one. And yeah, so basically that was my approach.

Transcripts of Users of GATE tool

The transcripts of the 3 users A, B and C of the GATE tool have been documented for analysis in the above experiments.

Transcription of User A during the concurrent verbal talk-aloud protocol for approximately 6 minutes

I'm opening it here and it's loads pretty quickly. So, that's fine. Not loading so quickly this time. All right, it's open. I have to resize the window a lot to get it to fit with other things. Not very smart about that. Uhh, Okay, Language Resource, New Gate Document, Okay, Umm. The screen with the URL is pretty confusing. There's many options on it. I'm gonna see about the South China Sea document now. It was like a task1.Umm, I think I have it. When you first load a document, it doesn't really come up. Okay, I have to double click it separately. Okay, I got that working. So, when I pull up the file, it does not have new annotations up. I click on annotation set and signal. Okay, so this file doesn't have that many signals. Umm, okay. I have an annotation here. This is not a proper annotation. Better leave these as signals. This one is a proper annotation. So now, I can load the annotation schema and it's hard to go up a level in the directory from the file selector. Okay, so I loaded the schema, this is co-specify-01.And, I need to get some other schemas. I have to pin it. And, there's one more signal on here. And, there's another one. Now, they are specified, Co-Specify-02. And, okay, I forgot the pinning. Okay. I guess, this one is up for it. Umm, so I have done all the annotations. So, I have done the annotations. I can keep going. Okay, I am reading through the paragraph. Umm, this file seems to have not many annotations. I can consider Beijing. I think there are not many more in this file. Umm, so I guess I am now ready to export. Okay, umm, very quick on the task file gate xml. I am not sure what inline xml is. I can say this task 1 binding and I save it. Okay, so that's done.

Transcription of User A during the retrospective report for approximately 2 minutes

Umm, so I opened the, I opened the tool, found the file that I'm gonna annotate. Then, I annotated it. There was not that much content to annotate. And I had to load the co-specify schemas, one by one. Then I exported it, when I was done under a new file.

Transcription of User B during the concurrent verbal talk-aloud protocol for approximately 6 minutes

Okay, so I'm going to start by clicking on the search tool on my laptop and search for GATE because I don't remember where I had it saved. Double clicking the folder, that's called GATE Developer 8.5.1 and I'm double clicking on the application within that folder. That's called GATE. Umm, while that's loading, I'm going to open another finder window. My computer just switched screens. Okay, I opened another finder window so that I can open the folder within which I have the task saved. So in finder on my Mac, I'm clicking on the folder to get to annotation project. Opening folder called task and within that folder is a folder called Documents Task1. Okay, I did it so that I just know where I am. Now, within GATE, it's still loading. So, I'm waiting, I'm waiting. Okay, it is open. Now, the GATE is open and I'm going to right click on Language Resources, Select new. I'm going to select annotation schema and click the book. Then I'm going to.re-find where this have been saved on my computer with this in Annotation project, task, schema groups, specify-01. I don't remember I can do more. Oh, I can, more than once. So schema-01's loaded. I'm clicking on the same language resources, new annotation schema, book. I get 02. Right clicking new annotation schema, book, specify-03. Okay, I'm gonna go for 3 for now. I hope, that's enough. I'm gonna right click on language resources, new, GATE document, move up a folder, Documents, task1. Okay, so I think I have everything I need. I'm going to double click on task1. It loads in the window. I click on annotation sets and then annotation sets, I scroll down. I'm looking for. Oh, no no. I'm clicking signal and then I'm scrolling down. I find some words and now, I'm gonna complete the annotation task. 'it', I believe, refers to strategic intent. So, going to call it, co-specify-01. I think, I have to hit this pin button. Nope, I don't. Okay. And then, I'm going to select the words, strategic intent and to also label that co-specify-01. So if they are linked, I need to double check by clicking off the check box of co-specify-01 and they both disappear. So I know I have done it correctly, I think. So, umm, I'm gonna read out. 'Strategic intent describes long term goals and aims rather than detached actions. It creates short term actions and goals connected to a larger picture. Given this'. So 'this', I'm gonna count as co-specify-02. I don't think it's just referring to strategic intent. I think it's referring to the fact that strategic intent describes long term goals. Umm, or maybe, it connects, it creates short term actions and goals connected to a larger picture. So, they've been this. So I'm gonna highlight ' It creates short term actions and goals connected to a larger picture.' as the concept that this is connecting to, I'm gonna label that co-specify-02. I'm gonna check the checkbox so that they disappear. They do. Okay, I'm gonna move on to the last highlighted section which is the last part of the next paragraph. So, 'some of this work'. So, all of that could be from the previous sentences. 'The other computing hypothesis presented above given supporting evidence. It is likely that China is putting in infrastructure in order to consolidate control over the area of natural resources, on sea and in air including fish stocks, oil and gas. Some of this work. I'm thinking work is putting in infrastructure. So, I'm going to make this co-specify-03 and then go back out of that and highlight putting. Umm, 'China is putting in'. I'm gonna highlight, putting in infrastructure and make that co-specify-03 as well. It's purple. I click it off and both of them disappear. So, then, I'm done with the annotation process. So, I have to now save the file. Umm, so, I'm going to go back and umm, right click on task1.xml to save as gate.xml and then rename it as task1_interview and save that in the same file. And then, I can just quit GATE and be done with it.

Transcription of User B during the retrospective report for approximately 2

minutes

So, to summarize, what I did in order to complete the annotation task is, first locate the GATE application on my computer. I opened the GATE application. While, the GATE application was loading, I located where the relevant files were on my computer so that I could find them easier, once GATE opened. I then, waited for GATE to open. Once, GATE fully loaded, I right clicked on language resources and multiple of the schema documents which are called specify-01, specify-02 and specify-03. Umm, then after I added those schema documents, I right clicked on language resources again and added the GATE document which is called task1.xml. Once, that loaded, I double clicked on it, so that I can see the text and then I checked the box on, umm, signal. I think is what's called to make sure that the targets lit up or highlighted. I scrolled down. I found the highlighted targets for each target, I read the text and labeled the target co-specify-01 and then labeled what the target was pointing to co-specify-01 as well by highlighting the relevant text, mousing over it and then waited for a little box to pop out and then choosing co-specify-01 from the dropdown or 02 or 03 whichever was relevant. After doing that for each of the three targets, the text, I, I saved as, the task, I saved the task as a file with a different name and then that completed the task.

Transcription of User C during the concurrent verbal talk-aloud protocol for approximately 6 minutes

So, I'm clicking on menu, I'm gonna type in GATE and I'll click on the GATE one with the image. And then, I'll go to language resources. I click annotation schema, the folder thing. And then, I'll find the documents folder and the folder button thing. Oh, I did that wrong! Okay, I have to double click on it. There you go. And then, Penn State, click. Sorry! Come on! And then Human Computer Interactions and the Research Methods and then annotation. I was in schema, so schema groups, Specify-01 and then ok. And then, I'll do that again. Okay, so new annotation schema, hold the button, Specify-02. Okay, maybe twice more. So, annotation schema, hold the button, Specify-03. Okay! New annotation schema, folder 4. And then I'd open a document which I think I'm gonna do here. Language Resources under GATE document, the folder that I'm gonna go to is Task and open Documents and then task1 and then that should be fine. Okay, open the documents and then annotation sets. And then put the signal on so that I can see it. And let's go down until I see the highlighted signal, I'm sorry, C1 and I need to see if it's appropriately signalled. So strategic intent describes something goals and it creates short term actions. Okay, so I think 'it' refers to a strategic intent. So, I'd say that's probably not what we are going for because it's a thing. So I'm gonna hit the green with the X on it meaning it is not appropriate. And, now go to the next one and I'd read that. So, 'given this', let's see what that's referring to. Associated content creates short term actions and goals connected to a larger picture, given this. So, I think that this means that the first creates short term actions and goals connected to the larger picture, essentially the whole previous thing. So I think that, that is correct. So, I'm gonna over around that and change it to Co-Specify-01 and I'll tag that down. And then, I will highlight what I think it was referring it to, so the whole previous sentence and hover over that until it comes up with only one that is Co-Specify-01. So, I" tag that down and move on to the next highlighted specification which is 'this work'. So some of 'this work' that refers to likely that China is putting in infrastructure in order to consolidate control over various natural resources on sea and in air including fish stocks, oil and gas. So, 'this work' is referring to China putting in infrastructure in order to everything else in that sentence. So, I think that, that's an action. So, that would be something that should probably, that is what we're looking for. So, I'll hover

over it. I really want to click on it. But that's not what I am supposed to do. And change that to Co-Specify-02 and tag that down and then I'll highlight the previous sentence and make that Co-Specify-02. So I've gotten that down, hover, waiting to do that correctly. Okay, so I think there we go, perfect! Tieing it down in the red pin. Give me a second to look maybe. And then, using one. There doesn't appear to be anymore co-specifications on this document. So, I will save it by going to the task document on the left. There you go and then Save as gate xml. I want to change the file name. So, I'm just gonna put my name and then practice for something and then save.

Transcription of User C during the retrospective report for approximately 2 minutes

So, I began by going to the menu and finding the GATE tool by typing in GATE and looking for the program and once, the program opened, I went to language resources, schema groups. To find it, I had to look through the documents to find the schema folder and then I opened 4 different schemas. And then, I went to language resources again. But, this time, I opened a GATE document and I went from the document folder with the task in it. I opened that and then I clicked on annotation sets and turned on the signals, so I could see what would be highlighted for me. Once, I found the signal, that was highlighted, I saw what that was and tried to find what it was referring to. I considered whether or not, it was something, that we were looking for or if it was a different reference that we didn't want. Once, I had decided, the first one was not what we wanted. So, I hit the green button with the X meaning that we've done what we wanted. And then, the next one that I found, that was highlighted was what we wanted. It was correct. So, I found what it was referring to and I labelled both of them as co-specification-01. And then, I did that for, I believe, the next one that we had but labelling them as co-specify-02. And then, once I was finished, then I saw there were no more words highlighted for me. I went to the document which is on the left and I went to Save As, titled it and then saved it.