

5-2011

Narrative analysis and computational model to predict interestingness of narratives

Laxman Thapa
University of Texas-Pan American

Follow this and additional works at: https://scholarworks.utrgv.edu/leg_etd



Part of the [Computer Sciences Commons](#)

Recommended Citation

Thapa, Laxman, "Narrative analysis and computational model to predict interestingness of narratives" (2011). *Theses and Dissertations - UTB/UTPA*. 81.
https://scholarworks.utrgv.edu/leg_etd/81

This Thesis is brought to you for free and open access by ScholarWorks @ UTRGV. It has been accepted for inclusion in Theses and Dissertations - UTB/UTPA by an authorized administrator of ScholarWorks @ UTRGV. For more information, please contact justin.white@utrgv.edu, william.flores01@utrgv.edu.

NARRATIVE ANALYSIS AND COMPUTATIONAL MODEL TO PREDICT
INTERESTINGNESS OF NARRATIVES

A Thesis
by
LAXMAN THAPA

Submitted to the Graduate School of
The University of Texas-Pan American
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2011

Major Subject: Computer Science

NARRATIVE ANALYSIS AND COMPUTATIONAL MODEL TO PREDICT
INTERESTINGNESS OF NARRATIVES

A Thesis
by
LAXMAN THAPA

COMMITTEE MEMBERS

Dr. Emmett Tomai
Chair of Committee

Dr. Richard Fowler
Committee Member

Dr. Laura Grabowski
Committee Member

May 2011

Copyright 2011 Laxman Thapa
All Right Reserved

ABSTRACT

Thapa, Laxman, Narrative Analysis and Computational Model to Predict Interestingness of Narratives. Master of Science (MS), May, 2011, 53 pp., 11 tables, 6 figures, 25 references.

In this research, I present results demonstrating the classification of the specially generated narratives by a machine agent by listening to human subject describing the same sets of the events. These classifications are based on human ratings of interestingness for many different recountings of the same stories. The classification is performed on various features selected after analyzing the different possible feature that affect on the interestingness of narratives. The features were extracted from the surface text as well as from annotations of how each narration relates to the content of the known story. I present the annotation process and resulting corpus, the feature selection, and experimental results for the task of predicting the interestingness.

DEDICATION

The completion of my master studies would not have been possible without the love and support of my family. My mother, father, brother, sister, sister in law, brother in law have continuously motivated me to pursue my career goal in every step of my life. They heartedly inspired, motivated and supported me in every aspect to accomplish this degree. I would like to dedicate my thesis to them.

ACKNOWLEDGMENTS

I am grateful to Dr. Emmett Tomai, my thesis supervisor and chair of the thesis committee, for all his mentoring and advice. His motivation and suggestion helped me a lot and I learned several things from him. His infinite patience and guidance is more than praiseworthy. He has supported me in every steps of this research work from the beginning. I would like to thank my thesis committee members: Prof. Dr. Richard Fowler, and Dr. Laura Grabowski to be part of it. Their advice, input, and comments on my dissertation helped to ensure the quality of my intellectual work.

I would also like to thank my friends with whom I discussed about my research work. Their generous suggestion is very helpful to pursue this research. I would like to thank all the professor at UTPA who teaches me that helps to complete my degree.

TABLE OF CONTENTS

	Page
ABSTRACT.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENTS.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER I. INTRODUCTION.....	1
Motivation.....	3
Thesis Organization.....	4
CHAPTER II. LITERATURE REVIEW	5
Narrative and Discourse Processing.....	5
Interestingness of Narratives.....	6
Rapport Corpus.....	7
Development Tools.....	8
CHAPTER III. WEB-BASED ANNOTATION TOOL.....	16
Introduction	16
Source Narrative Structure.....	18

Utterances.....	19
Annotation of Transcripts	20
CHAPTER IV. FEATURE IDENTIFICATION AND MINING.....	29
Number of Events.....	29
Number of Details.....	30
Number of Back Jumps.....	30
Un-annotated Utterances.....	31
Unique Words.....	31
Junk Words.....	32
Transcript Length.....	32
Non-selected Features.....	32
CHAPTER V. EXPERIMENT SETUP	34
First Phase Experiment.....	36
Second Phase Experiment.....	38
Accuracy Evaluation.....	39
CHAPTER VI. RESULT AND DISCUSSION.....	41
Result Analysis.....	46
Discussion and Future Work.....	48
REFERENCES.....	50
BIOGRAPHICAL SKETCH.....	53

LIST OF TABLES

	Page
Table 1: Individual ratings results with single Feature for Tweety dataset	41
Table 2: Individual ratings results for Tweety with binary feature set	42
Table 3: Individual ratings results for Sexual harassment with binary feature set	42
Table 4: Agreed ratings results for Tweety with binary feature set	43
Table 5: ratings results for Sexual harassment with binary feature set	43
Table 6: Individual ratings results for training on Tweety data and testing on Sexual harassment with binary feature set.....	43
Table 7: Comparing Individual ratings results for Tweety with Naïve Bayes binary feature and five-bin discretized data set.....	44
Table 8: Comparing Individual ratings results for Sexual harassment with Naïve Bayes binary feature and five-bin discretized data set.....	44
Table 9: Comparing agreed ratings results for Tweety with Naïve Bayes binary feature and five-bin discretized data set	45
Table 10: Comparing agreed ratings results for Sexual harassment with Naïve Bayes binary feature and five-bin discretized data set.....	45
Table 11: Comparing Individual ratings results for training on Tweety and testing on Sexual harassment with Naïve Bayes binary feature and five-bin discretized data set.....	45

LIST OF FIGURES

	Page
Figure 1: An excerpt from a transcript.....	19
Figure 2: Utterance structure of a transcript of tweety video clip	20
Figure 3: Event annotation example with event details.....	22
Figure 4: The web-based annotation tool interface for interestingness score	24
Figure 5: The web-based annotation tool interface for events mapping	25
Figure 6: Use of machine Learning Algorithm to classify input data	35

CHAPTER I

INTRODUCTION

Narrative is in form of a sequence of the related events that draws the attention of readers. Interest is a crucial factor in the motivation and memorability of narratives [7]. In face-to-face storytelling, a narrator has an opportunity to interact with the audience, and can gain their attention by mixing different activities in the telling [3]. Unlike face-to-face storytelling, written narratives must have intrinsic features that can generate cognitive interest to the story reader. Stories are found more or less interesting by different people at different times. However, despite this wide subjectivity, it is also true that some stories are much more consistently and generally interesting, while others are not. Interest is motivating and promotes memory, both of which are crucial issues for applications that seek to communicate via narrative [7].

What makes one narrative more interesting than another can be attributed to a wide variety of factors. The pattern of the story can bring interestingness to the narratives along with many other unidentified features. In general case, the writer has an imagined reader in mind. Suspense, surprise, curiosity and plot can generate interest in people [20]. Interest value of words, concept, abnormality, non-informative text, death, danger, power and sex can influence audience quality rating [20].

Interest also has a durational aspect. There are triggering factors that lead to continued interest over time. Structural position and structural organization aid in memorization and recall of events. Abnormal or novel information can induce a greater degree of interest, such as the

death of a 21-year old vs. the death of a 82-year old, or mundane vs. unusual situations. Much of the work in what makes a story interesting straddles the boundary between the content being presented and the method of presentation. Skillful presentation of the events, which may or may not correspond to the order and details of the actual events, can also be a factor in interest [4].

However there is no technology exists for judging how interesting a narrative is to different audiences. The judgment of the interestingness by reading large number of transcripts for any person is a tedious job. A rigorous effort is required to read and judge. Development of any computational model for prediction can reduce time. It requires deep analysis of the features which effect human cognitive behavior. Low-level textual features as well as high level structural features analysis are required for development of such system.

In this research, I present a work with a corpus that includes two sets of transcripts. Each contains over hundreds of different narrative discourses of the same story. By controlling the story being told, this corpus allows to investigate how discourse features and the telling of the story correlate to the judgments of interestingness. I describe corpus collections, and further annotation efforts I undertook with other team members to connect the surface text with features of the common underlying story content. The task of identifying more and less interesting versions of the story is characterized as a classification problem. I discuss the process of feature selection, then present several machine learning experiments over the corpus to train a model to predict interestingness. I have put an endeavor to identify the salient features which correlate with the interestingness of the narratives. The feature identification process went through different experimentation phases to check the correlation of the feature with the interestingness ratings. I have used widely used Machine Learning Algorithms such as Naïve Bayes Classifier

and Decision Tree Classification Algorithm for the development and testing of the model. The system can receive identified multiple features that influence the interestingness of narratives and also find the correlations of features with interestingness.

Motivation

Research on personal stories has largely been conducted in the field of psychology and the social sciences. These researches are typically conducted for the qualitative rather than quantitative analyses. The examples of related research on the accuracy of personal stories include the work of [20], particularly interestingness that leads to the productivity of storytelling.

There are less works found in the literature of computational linguistics that focus on human judgments of narrative interestingness. This would seem to be at least in part due to the diversity of possible factors that delve deeply into cognitive and situational modeling. It is difficult to isolate linguistic factors from personal context, and to separate differences in the *discourse*, the actual realization of the narrative, from differences in the underlying *story*, the sequence of events being related. There have been long-standing and on-going research and debate on what makes a good story [18]. The proposed work on analysis of interestingness of a narrative is slightly different, since it focuses primarily on how a story is told. The judgment of interestingness of a narrative is challenging whether it's done by a human or a computer. Most of the research on the quality of narrative has focused on fluency, relevance to topic and cohesion analysis; and less on the entertainment value. For example, Coh-Metrix, a computational tool developed by University of Memphis, can produce indices of the linguistic and discourse representations of a text and can be used to investigate the cohesion of the text [12]. Other

systems, such as Glosser, can mine student essays for relevance to a topic and coherence of the writing [24].

Thesis organization

Chapter I presents the introduction about the research, background and motivation. It also discusses the aspects of the interestingness of written narratives. Chapter II simply discusses the literatures that are used while conducting this research. Chapter II also talks about the development toolkits used for this research. It gives the overview of Natural Language Processing Toolkit (NLTK) and classifier used for classification of the narratives. Chapter III describes all the procedures to map the source transcripts of the narratives in more uniform that are used in analysis and development of the model. It discusses about the annotation of transcripts and guidelines used for annotation. Chapter IV describes the techniques used to extract the prominent features that correlate with Interestingness. It also discusses about the most prominent features used in the experiments and development. Chapter V exhibits the experiment setup and the hypothesis for the experimentation. It also describes the development of model and testing. Finally, chapter VI shows the outcome of the thesis and discussion of the result achieved from the thesis. It also discusses about the future works for this research work.

CHAPTER II

LITERATURE REVIEW

In this research, I have made effort to analyze the structural factors of the narratives such as events, details, unique words, junk words to identify how interesting is the story. The measure of Interestingness to a particular reader is somehow subjective rather than objective. The motive of this research is analysis of the present corpus and figuring out the structural features present in the narrative that can correlate with the interestingness. The literature about the work and process used to predict interestingness are explained in subsequent chapters.

Narrative and discourse processing

A Narrative is a story that is presented in various formats describing certain sets of the events sequences. Narratives are used in various fields for the dissemination of information.[13]

Narratives are found in many constructive formats such as written text, speech, songs, movies, television, theatre, photography etc. Narratives are referred as aspects of human psychology in many fields [13]. Narratives are widely used for the communication and illustration of ideas.

Narration is a process that imposes communication within the story among different event sequences. Narrative is one of the rhetorical modes of the discourse. Exposition, argumentation and description are other rhetorical mode of discourse. Narratives are presented in the way the audiences who are getting in different way can infer the meaning of the text of them. The way narratives are presented varies depending on the format, form and method it was presented.

Narratives form of those which are presented in television is different than those written in the newspaper. Same events are described in the way audiences can get the meaning of those. The

written texts are more descriptive than those presented with visual footage. Whatever the form, the content may concern real-world people and events.

Discourse Processing is a general approach to analyze written, spoken, signed language use or significant development technique to analyze the discourse [26]. Discourse Processing analyses several objects present in the discourse such as writing, talk, conversations, and communicative events. Discourse analysis is used in diversified field of study such as Psychology, Linguistic, Social Science, Cognitive Psychology, Human Geology etc. Each field applies different techniques to analyze the discourse and adjust the technique according to need of subject matter.

The source corpus of this research includes two sets of transcripts, each containing over a hundred different narrative discourses of the same story. The narratives are somehow different in structural format than other story models though it contains sequences of events as general story models. The transcripts of narratives are broken into several utterances and multiple lines. The transcripts are analyzed by using modern techniques and the event sequences present in the transcripts mapped to the master events of the source data. The detail of process of the event mapping of the narratives is discussed in Chapter 3.

Interestingness of Narratives

Interestingness of the narratives depends on many factors. There is no known measure to judge the interestingness of narratives. The interestingness of narrative is subjective and depends on the person who is judging it. . Structural position and structural organization aid in memorization and recall of events. Much of the work in what makes a story interesting straddles the boundary between the content being presented and the method of presentation. Skillful

presentation of the events, which may or may not correspond to the order and details of the actual events, can also be a factor in interest [4].

The corpus used for this research is quite different than others. As discussed earlier, the narrative is generated in form of transcripts. There is very little work in the field of Natural Language Processing (NLP) for judgment of interestingness. Whatever study has been done are in the field of Psychology and Social Science.

Rapport Corpus

This section gives the overview of the corpus used in this study. The corpus used in this study is the main source of the data. The section discusses how the corpus is generated. The corpus of transcripts of verbal narrative is resulted from the work of [11]. The corpus is named as *Rapport Corpus*. The project is involved in generating the transcripts by listening to several human subjects. The main motive of the project was to assess the potential of an animated virtual character (the Rapport Agent) to create more engagement and speech fluency, as compared to a real human listener. The Rapport agent was able to give non-verbal feedback to human speaker. The Rapport Agent tracks the real human speaker's prosody, head movements and body posture in real time, and rapidly generates timely feedback using head nods and postural mirroring.

During the experiment, several human subjects were first asked to watch two different video clips. The two video clips were different in nature. First video clip was old animated video clip from Warner Bros. having Tweety and Sylvester as main characters. The second video clip was the "CyberStalker" clip taken from a live-action segment from Edge Training Systems Inc.'s Sexual Harassment Awareness video. This video was named as *Sexual Harassment* video. Each

human subject was asked to tell the story of the videos to the virtual agents. Subjects were told that the virtual agent was an avatar of a real human who was listening to their stories.

Each subject used headset for the interaction with virtual agent while telling their story. Each story has been recorded by virtual agent. Different annotators later transcribed subjects' utterances from these recordings, and made annotations concerning their delivery. Those annotation of the recorded telling included codes for intonation, pause, pronunciation, laughter, and volume, from which researchers studied incomplete words, prolonged words, and pause fillers representing disfluencies in the subjects' storytelling.

The results demonstrated that the virtual agents' feedback immediacy elicited subjects' greater feelings of rapport. The overall duration of the subjects' verbal behaviors was longer when they interacted with a virtual agent that presented timely immediate feedback, as opposed to the other types of agents. Subjects talked longer when retelling the events of the second video that they viewed. However, the timely immediate feedback of the virtual agent did not facilitate improved recall performance.

The Rapport Corpus consists of 293 transcriptions of spoken narrative, 147 describing the events of the *Tweety* video and 146 describing the events of the *Sexual harassment* video. The transcripts were obtained from author of [11].

Development Tools

To pursue the goal of research, I along with my advisor used recent tools and technique used in these kind of Natural Language Processing related tasks. The tool used are discussed in the following subsections.

Natural Language Toolkit (NLTK)

Natural Language processing task is not confined to the limited fields. It comprises many research areas including computer science, linguistics, statistics, human computer interaction and many Artificial Intelligence related tasks. Any Natural Language Processing task required extensive knowledge of different fields. If we have better tool that process natural text easily, great amount of the time can be saved for the development and testing of any Natural computational model to be designed.

Natural Language Toolkit (NLTK) is a freely downloadable toolkit that includes extensive software, data, and documentation. NLTK is designed for four primary goal simplicity, consistency, extensibility and modularity. NLTK is written on Python Programming Language. Python is simply powerful programming language with excellent functionality for processing linguistic data. Python is heavily used in industry, scientific research, and education around the world. Python is used for productivity, quality and maintainability of software. NLTK provides many useful functions and library that can be easily used on the basis of requirement of the task. NLTK distribution is available for windows, Macintosh and Unix Platform.

NLTK is important for scientific, economic, social and cultural reason. NLTK is used in many areas within industry that includes people in human computer interaction, business information analysis, and web software development. More importantly, NLTK is used in the research where people work with humanities computing, corpus linguistic through to computer Science and Artificial Intelligences.

NLTK comes with rich library with several implemented parsers and chunkers. NLTK is best for the raw text processing and classification and information retrieval task. NLTK comes with many linguistic corpora that are analyzed and processed. It also includes scientific

computing library with support for multidimensional arrays and linear algebra, required for certain probability, tagging, clustering and classification task. NLTK also includes 2D plotting library for data visualization.

In my research work I have extensively used the NLTK implemented classifier for the development and testing my computational model. NLTK has implemented several Machine Learning classification algorithms present in the literature. The implementation of Naïve Bayes Classifier, Decision Tree Classifier and Maximum Entropy classifier are used in my research work for development.

Machine Learning Classifier

Naïve Bayes Classifier. Naïve Bayes Classifier is simple probabilistic classifier that classify based upon the Bayes theorem. The popularity of Naïve Bayes Classifier is that it can take input of the multiple independent features and apply probability for each feature and come up with the classifications [9]. Naïve Bayes Classifier is apparently simple to use and understand but it works very well in many problems. Naïve Bayes Classifier has outperformed many modern machine learning approaches to solve problems [19]. Naïve Bayes Classifier is also known as “independent feature model”. Naïve Bayes classifier assumes that there is absence of all other feature when it computes the probability of one feature. Every single feature contributes to the estimation and classification. To classify the input values, Naïve Bayes Classifier calculates the prior probability of each of the label from the training data by checking frequency of each label. Naïve Bayes. Naïve Bayes Classifier calculates the joint probability of all the features and classifies the label for the input features. The label is classified by likelihood estimation of each of the label. The label, whose likelihood estimation is higher, is then assigned to the input data.

Naïve classifier can be trained very efficiently for the supervised learning of features and classification.

Another way of understanding the naive Bayes classifier is that it chooses the most likely label for an input, under the assumption that every input value is generated by first choosing a class label for that input value, and then generating each feature, entirely independent of every other feature. Of course, this assumption is unrealistic; features are often highly dependent on one another. This simplifying assumption, known as the naive Bayes assumption (or independence assumption) makes it much easier to combine the contributions of the different features, since we don't need to worry about how they should interact with one another.

Decision Tree Classifier. Decision Tree Classifier work on the basis of Tree build from the feature set. Decision Tree model is easy to understand and develop and modify. Decision Tree is build on the basis of the set of the decision rule. Decision Tree Classifier Algorithm work for all the dataset varies from small to too big [16]. Decision Tree Classifier generates the output as binary tree like structure that give easy understanding of the computation of the classification. The number of the branches from the node varies according the possibility of the outcome of the events i.e. if we toss a coin, there is possibilities of outcomes are either head or tail. The leaves of the tree represent the classification label where as the internal node represents the branching criteria to go to the next depth. The classification of the unknown input data starts from the beginning of the root node. Every possibility is checked and the flow of the searching goes according the condition met in the internal node. The continuation of the branch selection is done until it reach to a leaf with some label.

The algorithm can learn the data from the training set of data and create a possible binary tree with a single root node having most prominent feature at it. Other internal node contains condition according the other features. The method of the making complete decision tree depends upon the rule associated to individual feature. Before learning the data from the training set, we must pick “decision stump” for the corpus. A decision stump is the process of classifying input data on the basis of the single feature. Each of the features should have certain decision stump for classification. The use of a feature in our decision tree depends on the performance of that feature when we classify the input data with respective decision stump. We build our decision by including the features based on their performance of classification using that feature. Once all the features have decision stump, the complete decision tree is generated.

Inter-rater Agreement. To ensure the high quality of annotation, multiple assessments of annotated data is necessary before using those data. Researchers are required to use hand-coded data that are labeled with categories. To develop and test a computational model, it is necessary to find the generated data are reliable or not. The fundamental assumption behind the different methodologies is that data are reliable if annotators can be shown to agree on the categories. If different raters produce consistently similar results, then it can be inferred that they have followed the similar rule and a similar understanding of the annotation guidelines and expectation from the annotators to perform consistently under this understanding. When there are fewer numbers of annotators, there can be issue of annotator bias. This issue can be solved if numbers of annotators are increased. The differences among annotations get smaller and smaller as the number of annotators grows. Increasing the number of annotators is the best strategy, because it reduces the chances of accidental personal biases [16]. The inter-rater agreement is

required for reliability and validity of the data. Inter-rater agreement or inter-annotator agreement is a degree of measurement of homogeneity or consensus of the rating among the annotator or judges who rated the certain things. It is very useful too in determining the range and deviation of measurement for a particular variable. If the agreement among the rater or annotators are diversely different or not in accordance then some new adjustment is necessary to measure, either need to change the scale of the measurement of the variable or some adjustment with the guidelines to the raters. The raters' disagreement for the rating may be the cause of the subject matter. The answer to the some task are subjective, in this case it's difficult to come with total agreement among annotators.

There are many statistical tools present in the literature to measure the inter-rater agreement among the raters. The some of the options are Cohen's' kappa, Fleiss Kappa, concordance correlation coefficient and inter rater correlation. In this research, Cohen's kappa and Fliess Kappa are used to determine the accordance of measurement among the annotator for event selection and interestingness rating.

Cohen's Kappa. Cohen's Kappa is the statistical measures of the inter-rater agreement among the raters [5,6]. Cohen's kappa depends upon the more robust methodology than just simple calculation of the percentage of agreement. Basically Cohen's Kappa measures the agreement of two raters who rates N numbers of items into C number of mutually exclusive categories. As discussed in previous sections Kappa calculation is must before using any data for further development. In our case, we applied Cohen's Kappa calculation for Interestingness of N numbers of transcripts and 5 categories of the interestingness rating score. Cohen's Kappa can be only used to measure the agreement among two annotators/raters. Cohen's kappa

continuously got popularity due to easy computation and its effectiveness in the classifier accuracy [21].

Cohen's Kappa Calculation. To verify the consistency of data, we have calculated the Cohen's kappa for two types of agreement, Even-level agreement and Interestingness-level agreement. This measure was inevitable for annotation agreement of annotator to proceed in the research. The Event-level agreement is to identify how well the annotators are agree on whether a particular event is described in a particular transcripts. The formula to calculate Cohens' Kappa as follow.

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

Where k is kappa ranging value from 0 to 1. Higher the value, greater the agreement between annotators. $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category.

Fleiss Kappa. As sanity check of the annotation agreement, we used another famous method used widely for measuring multiple-rater agreement. Fleiss Kappa is generalization of Scott's Pi statistic measure that defines inter-rater reliability. Fleiss Kappa is very effective when there are fixed number of raters assigning or categorizing a number of items. Fleiss Kappa gives best result when there is involvement of multiple raters in categorization [10]. It is used only with binary or nominal scale rating. The Kappa value (k) is measured between 0 and 1. Higher the value of k, more intensity of agreement. Fleiss Kappa got popularity than other statistical inter-rater reliability due to multiple rater facility than other. It assumes that although there are a fixed numbers of rater have rated the items, different items are rated by different individual. This feature is very novel than other statistical measure [21]. Many researchers have studies the

significance of Fleiss Kappa for the multiple rater reliability. Study shows that the number of categories and subject significantly affects the k value of the measurement. Fewer the categories, higher the Kappa [21].

Fleiss Kappa Calculation. Fleiss Kappa can be computed once we have information of number of raters, items, item ratings for each raters/annotators. The Kappa k can be defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

The numerator represents the degree of agreement that can be obtained above chance and denominator represents the degree of agreement that is actually achieved above the chance.

In our case, the dataset is rated by four different annotators. We calculated the kappa for Event-level agreement and Interestingness-level agreement. The Event-level agreement is calculated to know the agreement among annotator for inclusion / exclusion of events. Interestingness-level agreement was calculated for the agreement among raters for rating the transcript on the scale of 1 to 5.

CHAPTER III

WEB-BASED ANNOTATION TOOL

In this research work, the main source of the data is Rapport Corpus as explained in chapter II. I limited my experiment to the data sources created from the Rapport Corpus. The Rapport Corpus of narratives was identified from an existing corpus of narratives, collected as part of a different research effort. Gratch [11] has mentioned and described a series of experiments to study the rapport that people can develop with interactive animated virtual characters, which they refer to as Rapport Agent. The Rapport Corpus was created when multiple human subjects told their story to the Rapport agents. Each human subject watched two short videos clip and told the series of the events sequences to the Rapport Agent that they understand and remembered. Each transcript is broken into several utterances. The annotations of the transcripts are done by intensive effort of eight persons and annotation of the transcripts is done by four annotators individually for each of the transcript of two different video sources. Following subsection describes the process of annotation and the outcome annotations.

Introduction

Each transcript contains a series of event described by different human subject by watching two short videos. The structure of transcript was generated from the Rapport agent listening to human subject. The content of the transcript is not exactly as clean as text found in a general story or narratives. It contains many words as we can found in general stories along

with normal words, it also contains many special characters that denote different associated with them.

As discussed in the Rapport Corpus section, Rapport agent recorded and interacted with human subject during the story-telling. Each recorded transcript were annotated by different annotators to generated transcripts. For example ‘/’ symbols present in the transcript denotes micro pause (less than 150ms) where as ‘//’ specify pause between 150ms-1s. Symbol ‘///’ reflects pause of 1 second and more used by human subject during the process of storytelling. Following are list of symbols present in the transcripts that are generated from the transcription.

PAUSE

/	micro pause (less than 150 milliseconds)
//	pause (.150 - 1 second)
///	pause (1 second) +

PRONUNCIATION

()	(uncertain word) uncertain interpretation
(xxx)	unintelligible or untranscribable utterance
ex_a_c_tly	slower or emphasized
:	i li::ke it lengthening; the more colons, the more elongation
-	jona- incomplete word

VOLUME

CAPS	WATCH OUT	louder speech relative to the adjacent speech
* *	*oral sex*	softer speech or whisper

LAUGHTER

((laughter)) for a lot of laughter.

OTHER

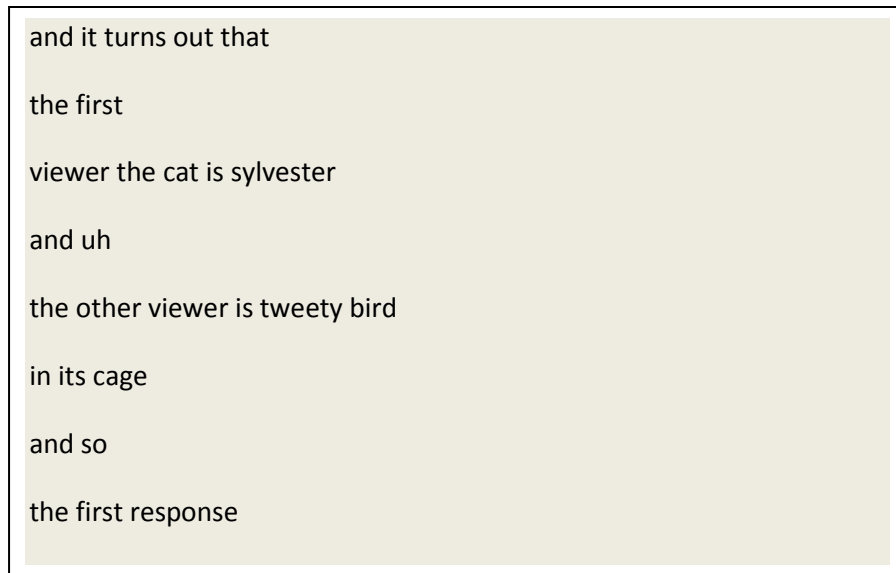
(()) ((knock on door interruption))
additional comments by transcriber

The above mentioned symbols are substantially present in the transcripts. We took out some symbols from the annotation that does not produce any significant effect on our research but kept some that may effect in the interestingness of the Narratives. For example, the words like “ex_a_c_tly”, “-“, “uh-“, “bow-ling” are present in significant amount and kept as it is for annotation of Interestingness. I discuss in detail about the transcript structure, utterances and annotation techniques we used for this research in the following subsections.

Source Narrative Structure

Narrative structure is the general framework that precisely exhibits the pattern and order narratives are presented to reader, listener and viewers. Narrative structure contains all the building block of narratives that bind readers to continue reading the narratives. Narrative structure contains scope and characters of narrative. It introduces all the basic situations and primary level of characterization exploring the characters background and characteristics events.

Unlike normal narrative structure, our resource transcripts might or might not present have followed the same pattern as it has been told by different human subjects by watching short video clips. The contents of the transcripts are the event sequences described by watching video clips. The events described were solely the human subjects’ own view towards the video clips. The transcripts of narrative do not follow any pattern of narrative structure. We are mostly interested in the event sequences descriptions. **Figure 1** exhibits the excerpts from one of the transcript with different utterances.



and it turns out that
the first
viewer the cat is sylvester
and uh
the other viewer is tweety bird
in its cage
and so
the first response

Figure 1: An excerpt from a transcript

Utterances

A source transcript contains several utterances. Each utterance comprises part of a sentence. Mostly, sentences are broken down into several utterances. Utterances contain the details and events sequence of the video clips. Sometimes more than two utterances should be combined to give the complete meaning of the sentence. **Figure 2** show a list of utterances from source transcripts

the uh cartoon animated cartoon movie begins
with a uh
a view camera view panning across a
a city skyline
and then focuses on a window
above which theres a uh sign that says
birdwatcher's association
then pans down and uh
there is a face a cat's face with binoculars
staring across at
an opposite building
an apartment building
and so you just see the uh the two eyes
staring through the binoculars
and

Figure 2: Utterance structure of a transcript of tweety video clip

As we can clearly see in the figure 2 that the utterances are not complete sentences. A single sentence is broken down into different several utterances with many exclamation words like “uh”, “um”, “ah” etc.

Annotation of Transcripts

Annotating the transcripts was the first tangible milestone of the research. Annotation of the transcripts was the principle task to reach to the point of predicting interestingness of narratives. Our transcripts were broken into multiple utterances. We could use different tactics to annotate the utterances in the way we could extract the information from the broken transcripts. Out of different possibilities, we decided to use the method in which annotation is done on the sentence level that can map all information to the more formal way. For better representation of

the event sequences, we created master event sets which refer each and every events sequence present in the video clips.

Master Event set

The two sets of the short videos contain many descriptions. For the better annotations of the transcript, master events set were necessary to map the event sequences presented in the transcript for accuracy of transcripts. It is necessary to figure out whether the statement in the transcripts accurately characterizes what happened in the video section. On the basis of interest of the research, it was the best to have focus on events (verbs) and link that refers to events in the video. We created master event set that includes all the possible events as well physically observable events in video.

The annotation contained a master list of events and details developed for each video. In the first phase, the development identified the overlapping sequences of directly observable physical events in each video. These are readily viewed and objective. These kinds of event also present all possible events description perceived by human being after watching it several times. But all of these events and actions were less likely to be mentioned by human narrator when they watch it couple of times. For example “the camera pan from left to right”, “the cat turned his head side by side”. The non-directly observable physical events such as sound playing on the background, the cat is on the third floor of the building were also included. . The second phase addressed the level of abstraction issue. A small development set of transcripts from each video's set of transcripts were taken as data source of the annotation. The events mentioned in these transcripts were also included in the master event list All of those events were added to master set that to make the master event set. The event mentioned in this development set were taken as

representative of entire set. The fine-grained events were then added to the new event as details as shown in **figure 3**.

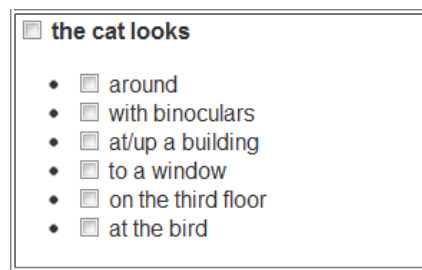


Figure 3: Event annotation example with event details

As also shown in figure 3, the master event list uses a semi-structured representation of events. The events are displayed in the way that makes annotator read and understand easily. Each event in the master is described as a SUBJECT-VERB phrase. Additional details are added as separate clauses, either direct object, prepositional phrase or directly quoted dialogue. This procedure was followed for both set of the transcript set.

Annotation Tool Interface

For each of the video, master event list is created that reflects the temporal sequence of events. The annotation references both event as a type and the part of the timeline in which it occurs. This fits well with the sequential nature of narrative. However, the videos also present situational details, such as where the characters are or what is visible behind them. These details are very less likely to be mentioned though the higher level of abstraction is presented in the event set. Therefore, these details are separated from the master event set. The details are not tied to any of the events in the master. The events presentation in the transcript have no boundary of the sequence, a narrator can describe any events in storytelling. Due to out of order event description, annotation was difficult for annotator.. With events, moving back in the master to a

previous event means that the narrator executed some form of flashback or sequencing error.

This is not true for situational details. To ameliorate this difficulty and confusion, the master list is divided into conceptually separate sections. The first shows the events, in sequence. The second shows situational details, which are not tied to that sequence. The third contains additional codes.

The interestingness scoring and events mapping may be affected by the degree of subjectivity due to different annotator. To overcome this, the annotation guide also contains codes to allow annotators to indicate utterances that do not strictly match the master event list. These codes were added to collect additional, more subjective data, but also to give the annotators various "none of the above" type options for difficult utterances. This was done to relieve pressure on the annotators to make things fit, helping keep the event and detail annotations simple and accurate. Different set of code set are created to include all the behavioral as well as physical characteristics. summary statements, such as story structure (e.g. "his first attempt..."), genre commentary (e.g. "every Tweety cartoon", "like all training videos") and emotional impressions about atmosphere, quality and value judgments are included in the first set of the codes where character assumptions such as appearance (e.g. race, physical characteristics), thoughts, motives, feelings and character traits are included in the second set of codes. The third deals with accuracy, allowing annotation of events and details missing from the master. The annotations for missing events and details are accurate, assumed, inaccurate and hypothetical. The final code is for observations made by the narrator that are not about the video at all (e.g. "I can't remember...").

The task of annotation was given to four undergraduate students from science and engineering background. They were hired for the summer specifically to complete this

annotation. Each annotators were assigned several training runs of the development set before annotating the real set of transcripts. The training runs were given to make them comfortable with the system of annotation. They would independently annotate two or three transcripts, then meet with a project leader to provide insight on the decisions they made and go over the guidelines again. The two of the annotator from the early training runs were also involved in the development of the *Tweety* master event list. The *Sexual harassment* master event list was done second and did not required same iterative development though it's very different and challenging due to heavy use of dialogue.

Transcript View: Rapport.SES.266.TS.SH.1.trs

okay the video i just watched was an old cartoon from the fifties or sixties
and it's been played hundreds and hundreds of times
it's the cat
um trying to
um capture the tweety bird
and
he runs up the
uh
water pipe or the uh um downspout pipe
for the um
roof of an apartment building
then he gets into the apartment
that tweety bird calls for his
her
um
caretaker which is the grandmother
she hits him over the head
he flies out of the building
um after she hits him on the head with an umbrella
then he um comes back up
and uh tries to get tweety bird again
and uh
the
tweety bird picks up a bowling ball and puts it down the downspout
down
down spout
pipe
and then
uh halfway down the building
the bowling ball ends up inside the cat
and the cat rolls out of the downspout
and into the bowling alley
yes

On a scale of 1-5, how well told (interesting, engaging, clear) is this narrative? Do not worry about how accurately it reflects the source video.

1 2 3 4 5

Figure 4: The web-based annotation tool interface for interestingness score.

Annotation Guideline

Each annotator was given a period of several weeks to complete each of the two transcript sets. They worked using a web interface that we developed for this task. For each set, they were instructed to first watch the source video. Then, they logged in to the tool where the transcripts were provided in random order (within each set). They were instructed to first randomly read 10 of the transcripts, and then work through the set one at a time. For each transcript, the annotator first went to a page where they read through the complete transcript in plain text, with one line per utterance and special annotation characters removed. Here they can rate the transcript for interestingness giving the score in the range of 1-5.

Transcript: Rapport.SES.333.NR.TS.SH.1.trs	Event master: tweety-master-v13-etomai.txt
Double click on an utterance to annotate	Select related details from master
<div style="border: 1px solid gray; padding: 5px;"><p>(17.589) and tweety bird was um (3/0/0)</p><p>(20.746) the little</p><p>(21.75) i beleive it was a canary</p><p>(24.185) and uh</p><p>(26.016) the cat was always after uh tweety bird (0/0/2)</p><p>(30.466) trying to attack it</p><p>(33.562) going after it's cage while it was in there but tweety (0/0/2)</p><p>(36.621) bird always out</p><p>(38.078) smar-</p><p>(39.351) sylvester</p><p>(41.636) and um</p><p>(43.119) the color was sharp</p><p>(46.349) there was uh a scene</p><p>(48.738) outside</p><p>(50.031) i guess in the house</p><p>(51.758) exterior vision</p><p>(56.064) with um</p><p>(58.381) hard to remember (0/0/1)</p><p>(61.007) with a canon</p><p>(62.281) tweety bi- i'm sorry</p><p>(64.014) sylvester shoots a cannon (0/0/2)</p><p>(65.931) like to blow up the bird and the bl-</p><p>(67.854) bird comes out of the cage but somehow uh (1/0/0)</p><p>(71.097) he's STILL ALIVE</p><p>(74.051) and shoots back at sylvester (0/0/2)</p><p>(77.097) so</p><p>(81.613) all in that both characters were still</p><p>(85.284) going strong</p><p>(88.106) yeah</p></div>	<div style="border: 1px solid gray; padding: 5px;"><p>Events Scene Details Codes</p><p>Scene 2</p><p><input type="checkbox"/> the bird is</p><ul style="list-style-type: none">• <input type="checkbox"/> swinging• <input type="checkbox"/> singing• <input type="checkbox"/> "When Irish Eyes are Smiling"<p><input type="checkbox"/> the cat climbs</p><ul style="list-style-type: none">• <input type="checkbox"/> up a drainpipe• <input type="checkbox"/> on the apartment building• <input type="checkbox"/> up to the bird<p><input type="checkbox"/> the cat waits</p><ul style="list-style-type: none">• <input type="checkbox"/> unnoticed• <input type="checkbox"/> next to the bird• <input type="checkbox"/> in the window• <input type="checkbox"/> swinging his finger/conducting with the singing<p><input type="checkbox"/> the bird looks at the cat</p></div>
<input type="button" value="Save"/>	

Figure 5: The web-based annotation tool interface for events mapping

The interestingness rated before the annotation of the event/detail mapping to give the annotator the look of general written narratives. Annotator read all the transcripts without the concept of the one by one event mapping to the masters' event set. The screenshot of the interface is shown in **figure 4**.

He or she then moved to the annotation page for that transcript. On this page, the transcript was displayed line by line on the left side. On the right side, the annotations were displayed as shown in **figure 5**.

They were instructed to go through the transcript line by line following these rules:

1. Double click on the utterance to select it (and save the last one)
2. If the utterance is information free (e.g. "um"), skip it
3. If the utterance is a connecting part of a bigger phrase (e.g. "where", "to", "at", "and then he"), and adds no details, skip it
4. Find and select the most appropriate events in the master that the utterance is describing
 - a. If there is no appropriate event, leave it blank
 - b. If you selected an event, select any details in that event that the utterance describes
 - c. The same events or details can be selected multiple times, people tend to say redundant things
5. Select any additional scene details given in the utterance that aren't covered by the events
6. Select any of the codes that pertain to the utterance

Data Storage

Total of 147 transcripts were in the *Tweety* data set, averaging 91 utterances per transcript. Five were used for development and not used for the final annotation. The remaining 142 were annotated by each of the 4 annotators. In average, 15 unique events were annotated out of 36 events present in the master event list. Pair-wise inter-rater agreement was calculated using Cohen's kappa [5, 6]. For the six pairings, the average kappa was 0.894 (std dev. 0.021).

The *Sexual harassment* data set consisted of 146 transcripts. It has average utterances per transcript were 94. Five were used in development and not used later for final annotation. The remaining 141 were annotated by the same 4 annotators. It has 38 master events in the master event list. An average of 18 unique events was annotated in each transcript. Pair wise inter-rater agreement was calculated for the six pairings has average kappa 0.769 (std dev 0.018).

For interestingness-level agreement, Fliess Kappa was calculated due to multiple annotator and rating categories more than two. The *Tweety* dataset has Kappa of 58.33 that is good enough to use this data for this kind of research. The kappa for Sexual Harassment dataset is 58.22.

A different relational database schema was created after post processing of the data to facilitate analysis. Most of the utterances are broken and do not directly represent grammatical or conceptual boundaries. Most of the events are span to multiple utterances in the transcripts. Many of those utterances are disfluencies, backtracking, or even unfinished sentences. The annotators were tasked only to identify the events and details being clearly described in any given utterance rather than attempt to annotate the span of each event. The annotation data was post-processed to convert the annotated events from the per-utterance link representation to an

utterance span representation for better analysis of event inclusion and ordering. Each event is considered to begin at the first utterance it is annotated, and end at the last utterance it is annotated. To account for redundant mentions, events are allowed to overlap only 1 utterance. Thus, if one event is annotated from utterance 1 to 10, and another at utterance 5, the first event is divided into two separate events.

CHAPTER IV

FEATURE IDENTIFICATION AND MINING

Samples from the *Tweety* dataset were analyzed to identify potential features that might correlate with interestingness ratings. For each potential feature, correlation tests were run with that feature in isolation, and small samples from the *Tweety* dataset with variation in that feature were analyzed. The *Sexual harassment* dataset was kept separate as a clean testing environment. Two classes of features were considered and selected. The first come from the content annotation and are intended to reflect the informational content of the narratives. The second come from the surface text of the transcripts and are intended to reflect the shallow discourse presentation.

Number of Events

The sequence of events in a narrative forms the backbone that moves the story forward. With the assumption that participants were unlikely to create totally new events out of thin air, the number of events from the video mentioned in the story provides a metric of the amount of content in the narrative. In preliminary tests, the number of events showed some positive correlation with judgments of high interestingness. It is possible that the annotators were influenced in their interestingness ratings by the accuracy of the transcripts, in which case the number of events might figure overly highly into the ratings. To combat this, the annotators rated each transcript prior to annotating it for content, and were explicitly told that interestingness was

not the same as accuracy. We also confirmed that the number of events was not in the top most informative features for any of the experimental conditions.

Number of Details

Similar to the number of events, the number of details comes from the content annotation. There is presence of many details that completes the events in the Narratives. Details are elements of the events that may or may not be mentioned by a narrator. Without details, it's hard to know the happening in the story. While events give some idea of overall content, the details indicate how complete the event descriptions are. Intuitively, I thought that detail density would be the most useful feature – normalizing against the number of words, utterances or events. However, preliminary experiments showed that the overall number of details was a more informative feature.

Number of Back Jumps

The third content feature is the number of times the narrator moves from a later event in the source video back to an earlier event, referred to as *back jumps*. This can be indicative of simple grammatical inversion (e.g. “He flew out the window when she clobbered him.”), but it can also indicate a deliberate flashback, or various forms of narrator error. Various features of event sequencing were investigated, but the only promising effect was found with the presence of numerous back jumps. Preliminary development testing shows that the number of back jump is also a crucial factor of determining the interestingness of narratives since most of the time it ranked in top 5 most informative features for the classification. Though the back jumps in the narratives found in limited quantity per transcript but comes out as one of the strong features for the Interestingness.

Un-annotated utterances

The fourth and final content feature is a simple percentage of the utterances in the transcript which were not annotated at all. This feature is intended to reflect the density of clearly communicated information. The set of codes in the annotation guide provided options for annotating extra events, details, commentary and even personal (narrator) information, so completely un-annotated utterances are likely to reflect broken sentences and filler that would make a story difficult to follow and readers' mind is diverted from the reading and leading it to low rating of the Narrative. In preliminary tests, the percentage of un-annotated utterances showed a negative correlation with interestingness ratings, with particularly low percentages showing higher ratings. While tested for those transcripts for which four annotators are agreed, it has been revealed that those transcripts contains very few un-annotated utterances than those transcripts for which annotators are not agreed on higher rating.

Unique words

Compelling narrative relies on the choice of words, and prior work has suggested that certain word types impact judgments of interestingness. *Unique words* in the surface text are identified as those words with a frequency of one over the entire corpus. The number of unique words reflects unusual word choices, which could be a positive or negative for readability and interestingness. Initial analysis showed that higher rated transcripts had less unique words, other features being the same. There is an error factor here with misspelled words in the transcription, as well as partial, broken words discussed below. The misspelled and broken words are accounted for a different feature. The unique words are calculated on the basis that those words are only present in that particular transcript out of entire corpus of the transcripts. The unique word calculation does not include stops word, broken words and junk words of the corpus.

Junk words

The original transcription of the Rapport corpus included many annotations of prosody. Almost all of these were removed in the textual presentation given to the annotators. However, some elements of the vocal presentation carry over to the textual presentation. We refer to these as *junk words*. First, there are the broken words, where the narrator began a word then stopped (e.g. “sylves-”). This was one of two prosodic annotations in the transcripts that was left in (the other was the use of all caps for loud emphasis). Second, there are the hesitation devices “um”, “er”, “ah”, etc. These were not specially annotated in the transcription, but are easily identified. The number of junk words in a transcript is intended to reflect negatively on the flow or smoothness of the narration. In preliminary tests, transcripts with similar features but fewer junk words tended to score higher interestingness ratings.

Transcript Length

The final feature we selected was simple transcript length, taken as word count. Given the annotation setup, it seemed possible that the annotators would be biased towards shorter transcripts that lessened their workload. However, analysis showed a positive correlation between notably long transcripts and higher interestingness ratings, even among transcripts with the same number of events. A longer transcript with the same events would seem to indicate less useful information density, but, in fact may indicate more skillful presentation of those events

Non-selected Features

Numerous other features were considered for inclusion, but no analysis or tests indicated any positive or negative correlations with interestingness ratings. The annotation codes, such as character motives, emotions, atmospheric descriptions, summary statements, inaccuracies and digressions, redundancy were too sparse in the data set to be used. It seems that these would be

good features for interestingness prediction, but collecting the data would be even more challenging than the event-oriented annotation task we completed. Lexical analysis was another promising area that yielded fewer useful features than expected. I experimented with the number of adjectives and adverbs used and words with positive and negative connotations before selecting the simpler unique and junk word features. Finally, I looked at additional higher level content features such as redundancy and causal chains, identified in related work [22], but only the back jumps feature was selected.

CHAPTER V

EXPERIMENT SETUP

The low overall agreement among annotators indicates the subjective nature of the rating task. Clearly this does not mean that there are no correlations to be found, but it does mean that there is no simple gold standard answer. Rather, I explore two assumptions here: first, that each individual annotator uses internally consistent guidelines, and second, that there are consistent guidelines across annotators. In each case, I am modeling these guidelines as correlation between the features we have selected and the interestingness ratings. The general hypothesis underlying all of these experiments is that a supervised learning algorithm can be trained using these features to predict interestingness ratings. Other hypotheses are that

- The interestingness rating can be predicted by single feature of the text.
- Individual annotators' interestingness ratings can be predicted from the complete set of features of the text.
- If annotators are agree on the interestingness rating of a transcript, these agreed ratings can be predicted from the complete set of features of the text.
- Interestingness rating can be predicted better with automatically descretized features.

In these experiments I have selected a Naïve Bayes (NB) classifier as a baseline supervised learning algorithm. To me and my advisor knowledge, there have been no prior interestingness experiments giving predictive accuracy using text features to compare against.

The NB algorithm is simple and well understood, and while the naïve assumption of feature independence certainly does not hold, it is known to perform as well or better than more sophisticated solutions [8]. Particularly in the field of text classification, there is ample precedent for this [25]. The NB also allows us to easily see which features in the training set are most informative. For these experiments, as I have already discussed in the Literature review section, I used the NLTK NB implementation. As a check for unknown implementation effects, I also ran the experiments on the NLTK Decision Tree (DT) algorithm. The training process and interestingness prediction process is shown in **figure 6**.

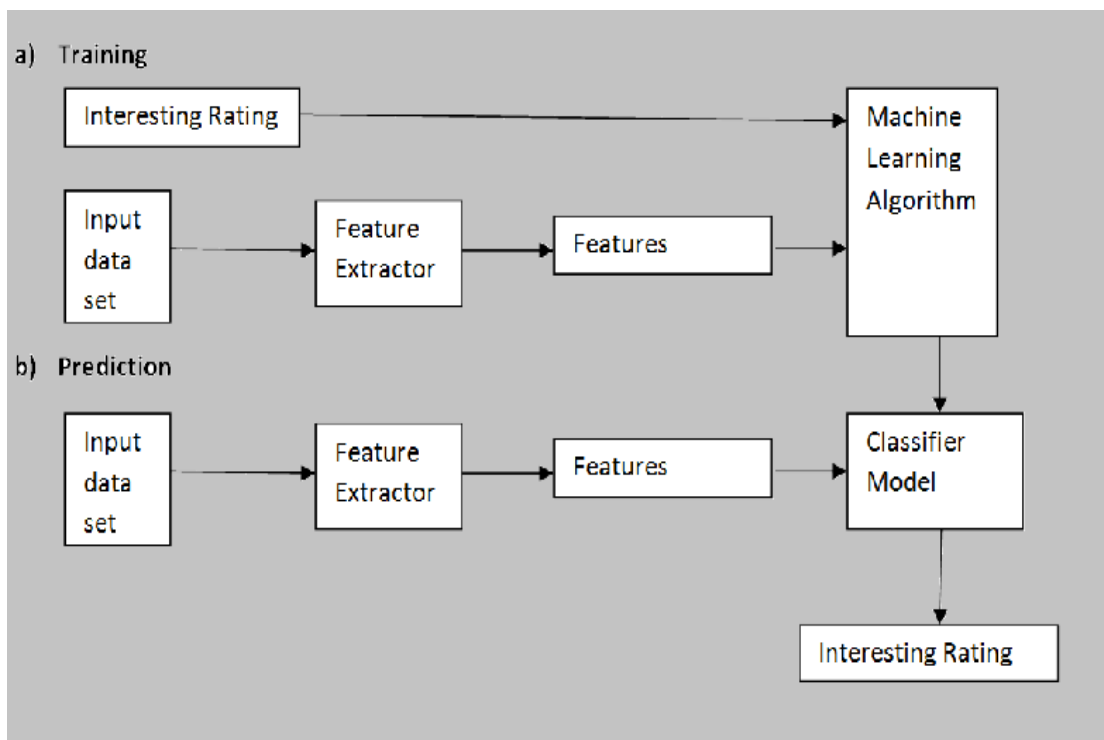


Figure 6: Use of machine Learning Algorithm to classify input data

There are two Phases of experiments, First phase experiment is done with Naïve binary feature discretization of continuous data where second phase of experiment was done by using multiple binning automatic discretization technique. Both phases of experiment contain similar

experiment setup with multiple features learning, only difference is discretization of continuous data. The first phase of experiment also includes the single feature experiment to explore the correlation of single feature with interestingness.

Each experiment tests the ability of the system to identify those transcripts which would be rated as more interesting. The transcripts which receive an interestingness rating of 4 or 5 from the relevant annotators are labeled as positive examples. The transcript with interestingness score with 1, 2 or 3 is labeled as negative examples. Due to small size of the data, 10-fold cross validation and multiple runs are used in all cases.

First Phase Experiment

The features are represented as binary distinctions. The number of events and the number of details are normalized against the total number of events and details annotated for a particular dataset. Those totals are calculated not from the annotation guide, but the union of events and details actually identified across the corpus. The number of back jumps, transcript length, the number of unique words and the number of junk words are normalized by their maximum value in any transcript in the corpus. The percentage of un-annotated utterances is already normalized. The features are discretized to binary distinctions. In most cases, a simple 50% threshold was used. In the case of transcript length, preliminary analysis indicated a positive correlation with notably long transcripts, so a higher threshold of 70% was used. In the case of un-annotated utterances, preliminary analysis indicated a positive correlation with notably low percentages, so a lower threshold of 30% was used.

The first set of experiment addresses the hypothesis that interestingness rating can be predicted by single feature of the text. NB and DT algorithms for each of the 4 annotators on

each of the 2 datasets were tested for this result. Among the multiple features, event inclusion feature is selected on top of other. The sequence of events in a narrative forms the backbone that moves the story forward. The number of events from the video mentioned in the story provides a metric of the amount of content in the narrative. I also experimented with other individual feature but event inclusion shows some positive result than others while combined together and analyzing the most informative features of test. I also combined the events inclusion features with other remaining feature individually but those experiment. These experiments are not listed due insignificant improvement than using single feature experiment. Naïve Bayes Classifier and Decision Tree Classifier are used for this task.

The second set of experiments address the hypothesis that individual annotators' interestingness ratings can be predicted from the complete set of features. This experiment is tested same as first with multiple features instead of single feature.

The third set of experiments address the hypothesis that if annotators are agree on the interestingness rating of a transcript, these agreed ratings can be predicted from the complete set of features. The agreement among the annotator is defined in term of k where value of k are varies on the basis of number of annotator agreement on transcript is positive or negative. If 4 annotators are agreed then k will be equivalent to 4. k will be 3 if only three annotators agree for the transcript. Only the subsets of transcripts with the specified agreement are used in each condition. For each of those transcripts, there are 4 examples, one for each annotator, where the annotator-specific features (number of events, details, back jumps and un-annotated utterances) may vary. Thus the training and testing is based on individual observations leading to unified conclusions. This experiment is tested for both dataset with multiple features.

As a follow-up, fourth experimental set was run to obtain results for training on the *Tweety* dataset and testing on the *Sexual harassment* dataset. Multiple features are used by prediction model to learn and predict. This was done for individual annotator ratings, as in the first experimental set.

Second Phase Experiment

As the experiment for the second phase of the testing, the same three set of experiment except single feature experiment were executed for five bin automatically discretized data. I experimented with variable binning of features like five, ten, and fifteen but there is no significant difference due to small range of continuous data. Later, it has been decided to use five bin discretization. The hypothesis of the testing was that interestingness rating can be predicted better with uniformly discretized data. The features are discretized by using equal width interval binning technique. [9] shows that there is negligible difference in accuracy to use this discretization technique than other more advanced techniques. This is perhaps the simplest and efficient method to automatically discretize the data. It is involved in the sorting the observed data and dividing them in k equally sized bins. The value of the k is supplied by user while development and fixed to a value where best results observed, In this experiment, the size of the bins is taken as five for discretizing each of the features described in the previous chapter. If a variable x is observed to have maximum value x_{max} and Minimum value is x_{min} then this method compute the bin width according to following formula.

$$\delta = \frac{x_{max} - x_{min}}{k}$$

The construction of the bin boundaries i.e. threshold is set at $x_{\min}+i\delta$, where $i=1,\dots,k-1$. This equal width bin technique creates k number of equal width bin. The continuous data then placed in the respective bin by checking the threshold of each bin from low to high.

Accuracy Evaluation

For each experimental condition, the overall accuracy of the NB and DT algorithms are generated for comparison. In all cases, the number of negative examples is greater than the number of positive examples, by nearly an order of magnitude in the most extreme case. Due to unbalanced classification of negative and positive examples for individual annotator, overall accuracy of NB and DT algorithm is meaningless. If model is judging 2 of 10 positive examples correctly from a dataset of 100 and guessing 80 of remaining 90 negative examples correctly, still the accuracy is 82%. Because this can significantly bias simple accuracy, for each condition the number of positive and negative examples along with the precision, recall and F-score are calculated for the NB algorithm. First phase of experiment is presented with the number of positive and negative examples along with the precision, recall and F-score where second phase of experiment presented only with precision, recall and F-score for comparison.

Precision and Recall are widely used for the evaluation of the most of the classification tasks. These two measures are simple metric that computes the fractions of the correct result by the system. Precision can be defined as the fraction of the retrieved positive result to the total retrieved result. In our context, precision is calculated using following formula.

$$precision = \frac{\text{retrieved correct positive result}}{\text{retrieved positive result}}$$

Similarly recall is defined as the retrieval of correct positive result from the total positive results in the testing set.

$$recall = \frac{\text{retrieved correct positive result}}{\text{total positive result in the test set}}$$

Precision and recall is used for evaluation of our system. The value of precision and recall show the correctness of the model for all the experiments.

Another statistics measure called F-score (F1) is also considered as accuracy of the system for the information retrieval task. F-score uses both precision and recall for computation. F1 score can be interpreted as weighted average of the precision and recall. The value of F1 ranges from 0 to 1. The accuracy is considered higher when the F1 score reaches to 1. The F1 score can be computed as follow.

$$F = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

CHAPTER VI

RESULT AND DISCUSSION

The first set of experiment is done with each feature individually. This experiment was conducted for number of events, number of details, back jumps, unique words, un-annotated utterances, junk words and transcript length. This experiment was performed for each of the annotators (consistently labeled as A, B, C and D). None of the individual feature is stand out alone to predict interestingness by its own. The result presented in **Table 1** with event inclusion feature on Naïve Bayes' binary discretization of continuous data. As explained previous section, event inclusion gives the overall view of narratives and other features work as supplement to events, so result with event inclusion is presented.

Annotator	Positive examples	Negative examples	NB Accuracy	DT Accuracy	P	R	F1
A	423	977	69.78	69.78	NA	NA	NA
B	397	1003	71.64	71.85	NA	NA	NA
C	612	788	68.92	68.92	63.80	66.0	64.89
D	208	1192	85.14	85.14	NA	NA	NA

Table 1: Individual ratings results for single Feature for Tweety dataset

Due to inefficiency of single feature prediction, all the features explained in the previous chapter are used for the testing and the result is much more improved than using single or combination of two features.

features. For each annotator, there were 142 total examples in the *Tweety* dataset, and 142 total examples in the *Sexual harassment* dataset. The result reflects the first phase of experiment with randomly discretized binary features. These results are shown in **Tables 2 and 3**.

Annotator	Positive examples	Negative examples	NB Accuracy	DT Accuracy	P	R	F1
A	43	99	80.5	81.2	74.4	78.1	76.2
B	40	102	81.2	80.5	84.0	52.5	64.6
C	62	80	70.1	70.1	85.5	88.3	86.9
D	21	121	87.0	85.7	72.2	93.4	81.5

Table 2: Individual ratings results for Tweety with binary feature set

Annotator	Positive example	Negative example	P	R	F1
A	55	87	80.4	67.3	73.3
B	17	125	72.7	47.1	57.1
C	54	88	84.1	68.5	75.5
D	16	126	75.0	56.3	64.3

Table 3: Individual ratings results for Sexual harassment with binary feature set

The third set of experiments were performed for $k = 3$ and $k = 4$, for each of the two datasets, for all feature inclusion conditions. The total number of examples varies by condition based on the agreement between annotators. There were no positive examples for $k = 4$ in the *Sexual harassment* dataset, thus the condition could not be run. These results are shown in **Tables 4 and 5**.

Agreement	Positive example	Negative example	P	R	F1
k=3	122	358	82.9	71.3	76.7
k=4	28	122	83.3	80.0	81.6

Table 4: Agreed ratings results for Tweety with binary feature set

Agreement	Positive example	Negative example	P	R	F1
k=3	82	366	76.7	56.1	64.8
k=4	N/A				

Table 5: Agreed ratings results for Sexual harassment with binary feature set

The last experiment for binary feature experiment per annotator was performed by training the system with *Tweety* dataset and testing on *Sexual Harassment* dataset with the ALL feature inclusion condition. These results are shown in **Table 6**.

Annotator	P	R	F1
A	58.0	52.7	55.2
B	27.0	58.8	37.0
C	49.1	51.9	50.5
D	27.8	62.5	38.5

Table 6: Individual ratings results for training on Tweety data and testing on Sexual harassment with binary feature set

Second phase of the experiment contains the result from the same experiment setup with 5-bin discretized data set. **Table 7 and 8** shows the result obtained by testing each annotator on *Tweety* dataset and *Sexual Harassment* dataset.

Annotator	Naïve Bayes binary feature			5-bin discretized feature		
	P	R	F1	P	R	F1
A	74.4	78.1	76.2	54.73	63.16	58.65
B	84.0	52.5	64.6	44.77	45.28	45.02
C	85.5	88.3	86.9	66.18	54.08	59.52
D	72.2	93.4	81.5	44.23	50.43	47.13

Table 7: Comparing Individual ratings results for Tweety with Naïve Bayes binary feature and five-bin discretized data set

Annotator	Naïve Bayes Binary Feature			5-bin discretized feature		
	P	R	F1	P	R	F1
A	80.4	67.3	73.3	56.88	58.22	57.54
B	72.7	47.1	57.1	19.23	10.81	13.84
C	84.1	68.5	75.5	50.09	45.47	47.67
D	75.0	56.3	64.3	33.88	47.67	39.61

Table 8: Comparing Individual ratings results for Sexual harassment with Naïve Bayes binary feature and five-bin discretized data set

Another experiment for 5-bin discretized feature set was performed on agreement data. This experiments were performed for $k = 3$ and $k = 4$, for each of the two datasets, for all feature inclusion conditions. The total number of examples varies by condition based on the agreement between annotators. There were no positive examples for $k = 4$ in the *Sexual harassment* dataset, thus the condition could not be run. These results are shown in **Tables 9 and 10**.

Agreement	Naïve Bayes Binary Feature			5-bin descretized feature		
	P	R	F1	P	R	F1
k=3	82.9	71.3	76.7	52.12	52.98	52.54
k=4	83.3	80.0	81.6	74.34	92.20	82.31

Table 9: comparing agreed ratings results for Tweety with Naïve Bayes binary feature and five-bin descretized data set

Agreement	Naïve Bayes Binary Feature			5-bin descretized feature		
	P	R	F1	P	R	F1
k=3	76.7	56.1	64.8	30.87	32.13	31.49
K=4	NA					

Table 10: Comparing agreed ratings results for Sexual harassment with Naïve Bayes binary feature and five-bin descretized data set

The final experiment for 5-bin descretized experiment per annotator was performed by training the system with *tweety* dataset and tested on *harassment* dataset with the ALL feature inclusion condition. These results are shown in **Table 11**.

Annotator	Naïve Bayes binary feature			5-bin descretized feature		
	P	R	F1	P	R	F1
A	58.0	52.7	55.2	57.69	54.55	56.07
B	27.0	58.8	37.0	43.48	58.82	50.00
C	49.1	51.9	50.5	52.72	53.70	53.21
D	27.8	62.5	38.5	44.44	50.00	47.06

Table 11: Comparing Individual ratings results for training on Tweety and testing on Sexual harassment with Naïve Bayes binary feature and five-bin descretized data set

Result Analysis

The prediction of the interestingness of transcript using single feature is not good. It guessed the interestingness all in negative way for all annotators except one. The single feature computational model predicted all the result as negative for those three annotators though it has accuracy with nearly 70% for all four annotators and above 85% for one annotator. The model was not able to predict positive example for three annotators. Due to no correct positive prediction, precision, recall and f1 score could not be computed. The experiment clearly shows that the single feature model cannot predict the interestingness of the narratives by its own.

The next set of results shown using simple NB algorithm, with naïve binary feature discretization, performed reasonably well in the prediction of individual annotators' interestingness ratings, based on the selected features. The overall accuracy of system with NB and DT algorithm is good without significant difference. Due to negligible difference between NB and DT overall accuracy, these measures are not presented in remaining experiments. NB algorithm is used henceforth for evaluations. The precision is above 70% in guessing the positive transcripts in both datasets. The strength of correlation with the selected features is quite annotator-dependent Recall in particular varies widely across annotators in the *Tweety* dataset, and drops below 50% in one case in the *Sexual harassment* dataset.

The third set of results shows that there is great chance of correlation of features with interestingness if the ratings are less subjective and consistent. The performance on the *Sexual harassment* dataset is again inferior.

The final set of experiment with binary feature discretization consist the result of cross dataset testing. The precision is suffered for two annotators and dropped to below 30s. It shows

that subjectivity of the research and the annotators' view on the different transcripts makes fluctuation in the result for precision and recall.

For the second phase of the experiment same set of features are used with five-bin automatic discretization technique. The results are presented with precision, recall and F1 score only due to insignificance of overall accuracy having large number of negative dataset. The first set of results show that the simple NB algorithm, with naïve 5-bin discretization, is not able to perform as well as binary feature dataset. The precision for per annotator dropped down to 44 % in the *Tweety* set and 19% percentage in the *Sexual Harassment* set. The strength of correlation with the selected features is quite annotator-dependent, also as expected. Recall in particular varies widely across annotators in the *Tweety* dataset. For three annotators the recall is more than 50% and mid 40s for an annotator. Similar result can be seen for *Sexual Harassment* data set where for one annotator the recall is dropped to 11%.

The second set of results with five-bin discretized dataset show that the same system is able to perform reasonably better than individual ratings on agreed dataset. The precision for $k=3$ agreement for tweety dataset is 52% where as 74% on $k=4$ agreement. The result obtained from the $k=4$ agreed data, the accuracy is more than 95% for both of the classifier where Recall is more than 92% and Precision is more than 74%. Again, performance on the *Sexual harassment* dataset is inferior due to no $k=4$ agreement among the annotators. The precision is also very low on $k=3$ agreed data.

The third set of result shows the cross-dataset testing. The recall for individual annotator testing is more than 50% for all annotators. Even though the subjectivity of the research and the

annotators' view on the different transcripts, has effect with fluctuation in the result, but 5-bin feature set vastly improved precision and recall.

Discussion and Future Work

The given source corpus is different in dialogue, event structure, and tone between a Sylvester and Tweety cartoon and a workplace sexual harassment training video. My effort to predict the interestingness by analyzing structural features has some positive sign when Naïve Bayes binary features are used. The threshold set by human for the discretization of the continuous features shows promising result for classification of interestingness for narratives. The result for the agreed transcripts has great result with all precision, recall and f-score above eighties for *Tweety* dataset on $k=4$ agreement. The 5-bin discretization of *Tweety* dataset work well, reaching recall of above 92%. However the experiment does not produce as good result as the first phase of the experiments for *Sexual Harassment* dataset where only $k=3$ agreement is present. This surely makes me to work in the future for figuring out known discretization technique that can be used for this research.

The feature set that was selected based on the *Tweety* dataset is shown to have predictive value in the *Sexual harassment* dataset as well. This suggests that the overall approach has merit beyond the development genre. However, the third experiment shows that the particular correlations trained from the one dataset do not transfer well to the other. This experiment is intended as a bridge and challenge to future work, included here only to provide some initial perspective on a larger generalization problem.

There are two clear directions for future work building on these results. First, the NB classifier is a popular and effective approach, but other classification algorithms should be

considered. In particular, the naïve binary discretization approach is a useful baseline, but various strategies for using continuous variables should be implemented to generate better result without using some absolute threshold for discretization. Further, more sophisticated models such as SVM [23] and CRF [16] have been used with success on text classification problems and should be compared. Second, feature selection in this set of experiments was not exhaustive either in the features considered or those selected. Given these results, it should be easier in the future to evaluate features in a more thorough way.

REFERENCES

- [1] Arie Ben-David (February 2008). "Comparison of classification accuracy using Cohen's Weighted Kappa". *Expert Systems with Applications: an International Journal* (Pergamon Press, Inc. Tarrytown, NY, USA) Vol. 34 No. 2 pp 825–832.
- [2] Artstein, R., and Poesio, M. (2008) Inter-Coder Agreement for Computational Linguistics, Association for Computational Linguistics, Vol. 34 No.4, pp 555-597.
- [3] Bavelas, J. L., Coates and Johnson, T. (2000). Listeners as Co-Narrators, *Journal of Personality and Social Psychology* Vol. 79 No.6, pp 941-952.
- [4] Brewer, W F. and Lichtenstein, E. H., A structural-effect theory of stories. 1982. *Journal of Pragmatics*, 6, pp 473-486.
- [5] Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* Vol. 20, pp 37-46.
- [6] Cohen J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* Vol. 70, pp 213-220
- [7] De Beaugrande, R. (1982). The story of grammars and the grammar of stories. *Journal of Pragmatics*. Vol. 6, pp383-422.
- [8] Domingos, P. and Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, Vol. 29, pp 103-130.
- [9] Dougherty, Kohavi, R. and Sahami, R.(1995) Supervised and unsupervised discretization of continuous features. *ICML*, pp 194–202.

- [10] Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382.
- [11] Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales M. and Morency, L. P. (2007), Can virtual humans be more engaging than real ones? 12th International Conference on Human-Computer Interaction, Beijing, China
- [12] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers* 36, pp 193-202.
- [13] Hevern, V. W. (2004). Introduction and general overview. *Narrative psychology: Internet and resource guide*.
- [14] Hidi, S. and Baird, W. (1986), Interestingness- A neglected variable in Discourse Processing, *Cognitive Science* Vol.10, pp 179-194.
- [15] John, G.H. and Langley, P.,(1995) Estimating Continuous Distributions in Bayesian Classifiers. *UAI*, pp 338–345.
- [16] Lafferty, J., McCallum, A. and Pereira, F.(2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Machine Learning, CONF* Vol. 18, pp 282-289.
- [17] Langley, P., Iba, W., and Thompson, K. (1992), An analysis of Bayesian classifier. In *proceedings of the Tenth National Conference on Artificial Intelligence, AAAI*, pp 399-306.
- [18] McCabe, A. and Peterson, C. (1984). What makes a good story?, *Journal of Psycholinguistic Research*, Vol 13 , No. 6, pp 457-480.
- [19] Rish, I., An empirical study of the Naïve Bayes Classifier.

- [20] Schank R.C. (1979), Interestingness: controlling inferences,. *Artificial Intelligence*, Vol. 12, pp 273-297.
- [21] Sim, J. and Wright, C. C. (2005) "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements" in *Physical Therapy*. Vol. 85, No. 3, pp. 257–268.
- [22] Tomai, E. Thapa, L., Gordan, L. and Kang S., Causality in Hundreds of Narratives of the Same Events.
- [23] Tong, S. and Koller, D. (2000) Support Vector Machine Active Learning with Applications to Text Classification. *Machine Learning. CONF* Vol. 17, pp 999-1006.
- [24] Villalón, J., Kearney, P., Calvo, R.A., Reimann, P. (2008) Glosser: Enhanced Feedback for Student Writing Tasks. In *Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies*.
- [25] Yang Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pp 412–420.
- [26] Zellig S., Harris (1952b.) *Discourse Analysis. The Structure of Language: Readings in the philosophy of language*.

BIOGRAPHICAL SKETCH

Laxman Thapa received MS in Computer Science from School of Engineering and Computer Science of University of Texas-Pan American. Before moving to United States, he received the Bachelor of the Science degree in Computer Science from the Tribhuvan University in 2003 from Nepal. He also received Master of Science degree in Computer Science and Information Technology from Central Department of Computer Science of Tribhuvan University in 2007. He has instructed several Computer Science and Information Technology subject from School level to the Undergraduate Level. He has also two years of experience of Industrial Software Development while working as Software Engineer for a Software Development Company. He has worked as Graduate Teaching Assistant and Research Assistant at University of Texas- Pan American. His Research interests are Natural Language Processing, Machine Learning, Computational Linguistic and Data mining.