A Bayesian Methodology for Semi-automated Task Analysis

Shu-Chiang Lin and Mark R. Lehto

School of Industrial Engineering, Purdue University, West Lafayette, IN 47906, USA
{slin,lehto}@purdue.edu

Abstract. This research proposes a new task analysis methodology that combines the fuzzy Bayesian model with classic task analysis methods to develop a semi-automated task analysis tool to better help traditional task analysts identify subtasks. We hypothesize that this approach could help task analysts identify activity units performed by the call center agent. The term activity units, in our study, represent the subtasks the agents perform during a remote troubleshooting process. We also investigate whether this tool could help predict the activity units as well. An effort-intensive field-based data collection for the call center's naturalistic decision making's environment was accomplished. A human expert and an additional 18 Purdue students participated in the validation of the assigned subtasks. The machine learning tool's performance was then examined. The preliminary results support our hypotheses that the fuzzy Bayesian based tool is able to learn and predict subtask categories from the agent/customer narrative telephone conversations.

Keywords: Task analysis, subtask, fuzzy Bayesian, machine learning, narratives.

1 Introduction

The purpose of this research is to propose a new task analysis methodology that combines a statistical approach, the fuzzy Bayesian model, with classic task analysis methods, to develop a semi-automated task analysis tool to better help traditional task analysts identify subtasks. We hypothesize that this approach could help task analysts identify activity units performed by the call center agent. The term activity units, in our study, represent the subtasks the agents perform during a remote troubleshooting process. We also investigate whether this tool could help predict the activity units as well [1].

1.1 Task Analysis

Task analysis is one of the oldest and most widely used methods in industrialized systems [2]. It has become a familiar tradition of practice for many human factors specialists. The essence of most task analysis techniques is a systematic identification and decomposition of a user's task into a number of small task components or activity units [3]. Although there are a wide range of available techniques to human factors specialists performing task analysis, identifying and decomposing a user's task into small task components remains as a well known difficult, impractically time-consuming, and expensive process that involves extensive manual effort [4].. Most of

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 697 – 704, 2007.

the task analysis methods also depend on how well a trained analyst understands the context in which he or she is analyzing the task; as a result, an inexperienced analyst will consume more time and more resources performing the analysis. Hence, there remains an essential and practical need for developing task analysis techniques to help practitioners perform task analysis efficiently and effectively.

1.2 Bayesian Methodology

There has been no formal research about utilizing a fuzzy Bayesian model based on semi-automated task analysis tool to help task analysts identify activity units. To learn why a fuzzy Bayesian method of particular is interest in developing an autonomous task analysis tool, we have reviewed some traditional knowledge acquisition strategies in terms of manipulating data. Studies have suggested statistical or probabilistic approaches for eliciting, filtering, parsing, and analyzing data or to classify narrative text into categories to provide decision support to troubleshooters because of their efficiency [5-7]. Among a number of these statistical approaches, applications of Bayesian model are proven to be successful in different disciplines. For example,; Hatakeyama et al. [8] show the Bayesian network model was effective to infer subjects' intention; Chatterjee [9] shows that the fuzzy Bayes model was superior in performance to that of the keyword model by classifying significant amount of free-text accident narratives that the keyword model had failed to classify; Zhou and Huang [10] verifies the efficiency and robustness of the Bayesian framework for hand tracking in high clutter background. Bayesian inference uses Bayes' Rule [11] to combine the various sources of evidence. Two Bayesian inferences are often proposed: classic Bayesian model and fuzzy Bayesian model. If multiple evidences are obtained, classic Bayesian assumes they are conditionally independent given that the hypothesis is true. On the contrary, fuzzy Bayesian makes strong dependence assumptions of the evidences. This approach reflects on the strongest evidence presented with no consideration given to negative evidence while co-occurrence of positive evidence is not aggregated. It is a more relaxed model that produces a relatively simple computation and conservative lower bound. Studies such as Zhu and Lehto [12] propose the Bayesian inference rule as a good statistical model to fit well the contingent relationship between index terms and words in the text. Other research such as Leman and Lehto [13], Qiu and Agogino [14], McCarthy [15], successfully apply fuzzy Bayesian model as a predictive technique to identify and/or predict print defect categories as well as help users diagnose print defects.

The fuzzy Bayesian model is an easily applied approach and could be implemented immediately by a task analyst with a laptop. We are convinced that application of the fuzzy Bayesian approach could lead to the implementation of more complete and systematic use of automated task analysis tools in the future.

2 Methods

Case studies are always vital and have contributed to both theoretical and empirical research in the naturalistic decision making (NDM) environment [16-17]. In this research, we utilize a conventional task analysis tool by recording traditional one-on-one phone conversations at the call center where the customers call to report various printer problems and the knowledge agents help troubleshoot the problems on the phone. The experimental process is conducted in four phases data collection, data manipulation, machine learning environment, and tool evaluation [1].

2.1 Data Collection and Manipulation

With the help of the Remote Print Defect Diagnostics research team, an effort-intensive field-based approach for the call center's naturalistic decision making's environment was accomplished. There were over 120 customer calls collected onsite. 24 experienced agents were observed during their normal working hours.. They were all 25 years of age or older and had at least 4 months or longer and had at least 1500 calls troubleshooting experience. Prior to the data collection, all the knowledge agents were voluntarily participated well informed the purpose of our research and our data collection process, which were the same as what the knowledge agents' mentors and/or their supervisors were constantly doing - monitoring the knowledge agents on the phone on a daily basis to help improve their remote troubleshooting skills as well as communication abilities. More than 60 hours of video/audio data with a total of 150 calls were collected.

In Phase 2 data manipulation, the knowledge agents' daily routine troubleshooting tasks were documented, the recorded audio data were transcribed into 770 pages of text with about 179,000 words, and a task decomposition table for agent with 72 subtasks was created. These subtasks were then tagged to the agent-customer conversation narratives to train and test the machine learning tool.

A significant amount of time and effort was spent in defining and assigning the subtask categories as well as in validating their reliability. There were a total of 20 people, including 2 human experts and 18 Purdue senior or junior students, participated in a four-step validation process to cross validate the assigned subtask categories. All of the students were eighteen years of age or older and had experiences using printers. They all had task analysis background. The author discussed constantly with the participants in validation process the discrepancies and made revisions to the assigned subtasks accordingly. The validation took approximately a year to complete.

2.2 Machine Learning Environment

In our study, we use the fuzzy Bayesian model to classify subtasks into categories using the expression:

$$P(S_i|E) = MAX_j \frac{P(E_j|S_i)P(S_i)}{P(E_j)} \tag{1}$$

where $P(S_i|E)$ is the posterior probability of subtask S_i is true given the evidence E is present, $P(E_j|S_i)$ is the probability of obtaining the evidence E_j given that the subtask S_i is true, $P(S_i)$ is the prior probability of the subtask being true prior to obtaining the evidence E_j, and "MAX is used to assign the maximum value of calculated $P(E_j|S_i)*P(S_i)/P(E_j)$. In our study, we consider the words used by the agent and customer as the sources of evidence, and describe what the agent said when performing subtask A_i by a word vector $WAi = (WAi_1, WAi_2, \dots ,WAi_q)$ of length q, where WAi_1, WAi_2, … ,WAi_q are the q words in the dialog. Similarly, we describe what the customer said in response to the agent's question by a second word vector WCi = (WCi_1, WCi_2, … ,WCi_q) of length q, where WCi_1, WCi_2, … ,WCi_q are the q words in the dialog. We also consider that A_i is potentially relevant to WAi, WCi-1, and WCi for i greater than 1.

Following this approach we derive equation (2) from equation (1) to calculate the posterior probability of subtask A_i:

$$\begin{aligned}&P(Ai|WAi,WCi,WCi\text{-}1)=MAX[P(WAi|Ai)P(Ai)/P(WAi),\ P(WCi|Ai)P(Ai)/P(WCi),\\&\qquad P(WCi\text{-}1|Ai)P(Ai)/P(WCi\text{-}1)]\\&=MAX\ \{MAX_j\ [P(WAi_j|Ai)*P(Ai)/P(WAi_j)],\ MAX_j\ [P(WCi_j|Ai)\ *P(Ai)/P(WCi_j)],\\&\qquad MAX_j\ [P(WC(i\text{-}1)_j|Ai)*P(Ai)/P(WC(i\text{-}1)_j)]\}\quad \text{for } j=1,2,\ldots.q\end{aligned}\tag{2}$$

In practice, we implement Textminer program [1] as our machine learning tool to carry out the above calculations. A snapshot of the Textminer interface is shown in Figure 1.

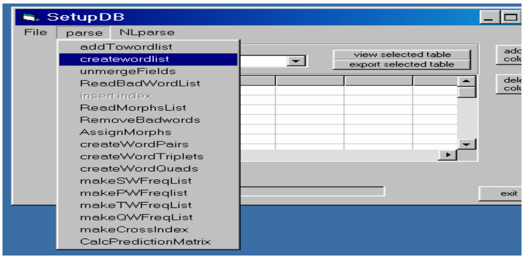

Fig. 1. A snapshot of the machine learning tool – Textminer*

The inputs of the Textminer learning tool are the words said by the agent and the customer. The conversation narratives were parsed and the keywords were elicited. The combinations of words appearing in the narratives were than used as candidates for subtask category predictors. Two-thirds of the 5184 narratives were randomly assigned to the training data set and the other one-third of the narratives was used as testing set. This process was repeated 10 times to allow for cross-validation of the accuracy of the model predictions. The output of each run was the conditional probability of subtask S_i performed by the agent. The preliminary tool performance was evaluated by hit rate and false alarm rates.

3 Results and Discussion

Extensive calculations were carried out and substantial amounts of quantitative and qualitative results, including over 400 tables with contents ranging from 41 to 4.3 million records, were generated during the model development process. The preliminary results summarized below have supported the hypothesis that the tool developed is able to learn and predict subtask categories from the telephone conversation between the customers and the human agents.

* All figures and tables illustrated in this paper are cited from reference [1].

3.1 Model Predictions

A total of thirteen prediction results were obtained and recorded in thirteen prediction tables respectively for fuzzy Bayes model. These tables, each with 5184 prediction results, predicted the subtask categories for various word combinations, ranging from single word alone to single_pair_three_four combination. Table 1 illustrates an example of the prediction table for combinations of single_pair_three_four words. Note that each record reveals the predicted subtask category, the keyword(s) used to make the prediction, and the strength of the subtask category for the particular predictor.

Table 1. A sample list of fuzzy Bayes prediction results based on a combination of single_pair_three_four words predictions[*]

ID[1]	Category[2]	Prediction[3]	PredAcc[4]	Strength[5]	Predictor[6]
2	211	211	1	1.0000	firstname&your
6	251	251	1	0.8333	Model&of&printer
7	627	626	0	0.6154	back&I'll&right
9	253	253	1	0.9091	on&one&serialnumber
10	254	371	0	0.1149	Right
11	252	621	0	0.3333	Its&said
14	541	541	1	1.0000	bye&day
17	611	521	0	0.5556	csonumber

[1] ID: Identification of a dialog/narrative.

[2] Category: Pre-assigned subtask category.

[3] Prediction: The predicted subtask category based on a combination of words.

[4] PredAcc: Prediction accuracy for that dialog/narrative. 1 means correct hit, 0 if incorrect hit.

[5] Strength: The probability of the subtask category for the particular Predictor.

[6] Predictor: The keyword(s) used to make the prediction.

A contingency table listing three most important and three least important predictors from each subtask category was also derived to illustrate the relationships between the word/word combinations and subtask categories. The word/word combinations are stratified based on subtask categories, strength of the predictions, and types of word combinations. Table 2 illustrates an example of one subtask with a portion of the word combinations.

3.2 Model Performance

The overall hit rates and number of times subtasks were correctly predicted are given in Table 3. Separate values also are given for each of the 10 cross-validation runs. The predicted results are further broken down into those for the training and testing sets. Some variation in the number of hits among the different runs was present, but overall the results are remarkably consistent, as illustrated by Figure 2.

[*] All figures and tables illustrated in this paper are cited from reference [1].

Table 2. Partial listing of a contingency table for one subtask and word/word combinations*

A[1]	S/W[2]	Single Word[3]	SF[4]	SP[5]	Pair Words[6]	PF[7]	PP[8]
321	S	EIGHTZEROZEROnumber	9	1.0000	EIGHTZEROZEROnumber&Okay	7	1.0000
321	S	dollarTWOPOINTFIVE	6	1.0000	charge&there	6	1.0000
321	W	on	10	0.0085	of&Okay	6	0.0165
321	W	Right	5	0.0084	Okay&on	7	0.0140

[1] A: Pre-assigned subtask category.

[2] S/W: Strong/Weak. Strong means that word or word combination is among the top three predictors with highest conditional probability of the subtask category. Weak means that word or word combination is among the bottom three predictors.

[3] Single Word: Parsed single word from single-word frequency list.

[4] SF: Single word frequency. Number of times that single word was assigned that subtask category.

[5] SP: Strength/conditional probability of that single word for the subtask category.

[6] Pair Words: Parsed two-word combination from two-word frequency list.

[7] PF: Pair words frequency. Number of times that two-word combination was assigned that subtask category.

[8] PP: Strength/conditional probability of that two-word combination for the subtask category.

Table 3. Number of hits based on 10 randomized cross-validation runs*

Run	1	2	3	4	5	6	7	8	9	10	Mean	Std.Dev	Hit
All data	2931	2784	2791	2809	2798	2830	2859	2758	2891	2867	2831.8	54.15	0.5464
Training	2212	2085	2102	2090	2039	2130	2159	2068	2170	2146	2120.2	52.60	0.6106
Testing	719	699	689	719	758	701	701	689	720	722	711.8	20.68	0.4160

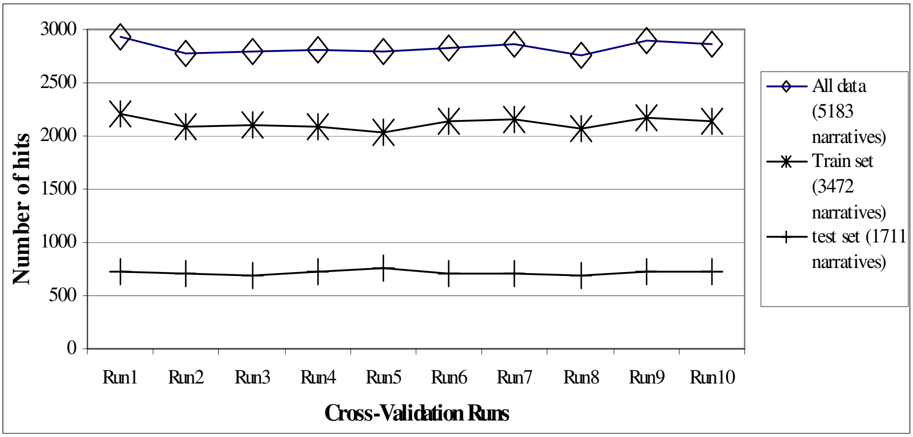

Fig. 2. Number of hits based on 10 randomized cross-validation runs*

* All figures and tables illustrated in this paper are cited from reference [1].

Table 4. ANOVA table to test effect of run on number of hits*

Source of vari.	Sum of squares	DF	Mean squares	F	P-value	F crit
Between Runs	0.0024	9	0.0003	0.0269	0.9999	2.3928
Within Runs	0.1987	20	0.010			
Total	0.2011	29				

ANOVA analysis was conducted to test whether differences between the ten runs statistically significant. The results of this analysis are given in Table 4, and confirm that the differences among the ten runs were not statistically significant (p = 0.9999).

The statistical result reveals the tool's robust prediction for our dataset, and supports the conclusion that we can randomly pick up two-thirds of the dataset as a training set and the remaining one-third as a testing dataset and run the tool one time instead of ten times while achieving similar predictions without too much variation. Student's paired t-test of the average hit rates across data sets shows the tool performs more accurately for training set than for testing set ($p=2.409x10^{-10}$ with a two-tailed distribution).

Since the average hit rate 56.55% is significantly greater than the false alarm rate, we conclude that our hypotheses are supported and that the tool is able to learn or predict subtask categories from the agent's and the customer's telephone conversations.

4 Promising Future Studies and Applications

The results presented here serve as a starting point and resource that enables future studies of applications of Bayesian theories to the very important research area of task analysis. With the abundance of both quantitative and qualitative data and results obtained in this work, many advanced analyses such as analyses on coarse subtask categories are being carried out now. The promising results have also provided the resources to explore exciting applications of hybrid Bayes models to many other areas of the naturalistic decision making environment.

References

1. Lin, S.: Machine A Fuzzy Bayesian Model Based Semi-Automated Task Analysis. Unpublished Ph.D. Dissertation. Purdue University, West Lafayette, IN (2006)
2. Luczak, H., Kabel, T., Licht, T.: Task Design and Motivation. In: Salvendy, G. (ed.) Handbook of Human Factors and ergonomics, 3rd edn., Ch. 15, pp. 384–427. Wiley, Hoboken, NJ (2006)
3. Morgeson, F.P., Medsker, G.J., Campion, M.A.: Job and team design. In: Salvendy, G. (ed.) Handbook of Human Factors and ergonomics, 3rd edn., Ch. 16, pp. 428–457. Wiley, Hoboken, NJ (2006)
4. Jeffries, R.: The role of task analysis in design of software. In: Helander, M., Landauer, T.K., Prabhu, P.V. (eds.) Handbook of human-computer interaction, 2nd edn., pp. 347–359. Elsevier Science, Amsterdam (1997)

* All figures and tables illustrated in this paper are cited from reference [1].

5. Yamamoto, H., Sagisaka, Y.: Multi-Class Composite N-gram based on connectiondirection. In: Proceedings IEEE International Conference on Acoustics, Speech & Signal Processing, vol. 1, pp. 533–536 (1999)
6. Salton, G., McGill, M.J.: Introduction to Modem Information Retrieval. McGraw-Hill, New York (1983)
7. Bookstein, A.: Probability and fuzzy-set applications to information retrieval. Annual review conformation science and technology 20, 117–151 (1985)
8. Hatakeyama, N., Furuta, K., Nakata, K.: Model of Intention Inference Using Bayesian Network. In: Stephanidis, C., Jacko, J.A. (eds.) Human-Centred Computing (v2). Human-computer interaction: proceedings of HCI International 2003, pp. 390–394. Lawrence Erlbaum, Mahwah, NJ (2003)
9. Chatterjee, S.: A connectionist approach for classifying accident narratives. Unpublished Ph.D. Dissertation. Purdue University, West Lafayette, IN (1998)
10. Zhou, H., Huang, T.S.: A Bayesian Framework for Real-Time 3D Hand Tracking in High Clutter Background. In: Jacko, J.A., Stephanidis, C. (eds.) Human-Centred Computing (v1). Human-computer interaction: proceedings of HCI International 2003, pp. 1303–1307. Lawrence Erlbaum, Mahwah, NJ (2003)
11. Savage, L.J.: The Foundations of Statistics. Wiley, New York (1954)
12. Zhu, W., Lehto, M.R.: Decision support for indexing and retrieval of information in hypertext system. International Journal of Human Computer Interaction 11, 349–371 (1999)
13. Leman, S., Lehto, M.R.: Interactive decision support system to predict print quality. Ergonomics 46(1-3), 52–67 (2003)
14. Qiu, S., Agogino, A.M.: A Fusion of Bayesian and Fuzzy Analysis for Print Faults Diagnosis. In: Proceedings of The International Society for Computers and Their Application-ISCA 16th International Conference, pp. 229–232 (2001)
15. McCarthy, P.: Machine Learning Applications For Pattern Recognition Within Call Center Data. Unpublished master thesis. Purdue University, West Lafayette, IN (2002)
16. Zsambok, C.E.: Naturalistic Decision Making: Where Are We Now. In: Zsambok, C.E., Klein, G.A. (eds.) Naturalistic decision making, Ch. 1, pp. 3–16. Lawrence Erlbaum, Mahwah, NJ (1997)
17. Hutton, R.J.B., Miller, T.E., Thordsen, M.L.: Decision-Centered Design: Leveraging Cognitive Task Analysis in Design. In: Hollnagel, E. (ed.) Handbook Of Cognitive Task Design, pp. 383–416. Erlbaum, Mahwah, NJ (2003)