Evaluating Improvement in Situation Awareness and Decision-Making Through Automation

Erin Zaroukian
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, USA
erin.g.zaroukian.civ@mail.mil

Justine Caylor
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, USA
justine.p.caylor.ctr@mail.mil

Michelle Vanni
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Adelphi, MD, USA
michelle.t.vanni.civ@mail.mil

Sue Kase
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, USA
sue.e.kase.civ@mail.mil

***Abstract*— Automated systems such as information extraction (IE) pipelines are designed to facilitate Situation Awareness by providing human decision makers with relevant information, but beyond the validity of the pipeline itself, designing the output of the pipeline for optimal human understanding should be a goal. This paper presents results comparing comprehension of text documents with and without markup from a (simulated) IE pipeline. While a previous paper suggests that markup hurts both objective and subjective measures of performance, this paper uses hand-generated markup designed to be maximally accurate and task relevant, finding more favorable results. These results, however, still point toward the limitations of markup and the importance of the task it is intended to facilitate.**

Keywords—information extraction, situation awareness, deductive reasoning, visual search, workload, usability, automation, decision making

I. Introduction

Good intelligence is critical for maintaining Situation Awareness (SA), but finding resources to tackle large amounts of intelligence documents is a widespread bottleneck. To address big data situations such as this, Information Extraction (IE) pipelines have been developed to automatically extract relevant information from large collections of resources, and the output of these pipelines, especially during development, is often presented to human analysts as markup on text input. Research on these pipelines typically focuses on precision and recall of the outputs, and while these measures are important for extracting correct information, the actual usefulness of this information to the end (human) user is often overlooked. Work that has focused on human users typically looks at user experience with highly specific, low-level features such as font size, color, and serifs, e.g., [1-4]. Little research is available on complex markup and its effect, both in terms of the information chosen for markup and the way in which it is marked up, on reading comprehension. In particular, while many different markup schemes are in use, there is a paucity of work comparing text with markup to text without markup to assess the value added by markup.

This paper asks whether markup actually improves human comprehension of text documents. Comprehension is measured objectively as the accuracy and speed with which participants answer questions about the text, and it is measured subjectively through ratings of workload and preference. In a previous paper [5], we asked this question using a pre-existing IE pipeline [6-7] and found, somewhat surprisingly, that marked-up text leads to worse comprehension (lower accuracy, slower response times, higher workload ratings, and lower preference ratings) than comparable text without markup. This paper presents a follow up experiment in which we attempt to stack the cards in favor of markup, using hand-generated markup that aims to be as accurate and task-relevant as possible. Nonetheless, find little evidence that participants perform better with markup. We do, however, find suggestions that participants generally preferred to have markup present. Further investigation may shed more light on why this pattern emerged and what it means for creating useful markup for improved SA.

This research was supported in part by an appointment to the Student Research Participation Program at the U.S. Army Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between U.S. Department of Energy and USARL.

II. Previous Experiment

A previous experiment [5] was conducted to compare human comprehension of simple text documents with and without markup from an existing IE pipeline to determine whether this markup improves human comprehension of these text documents. Comprehension was measured objectively as the accuracy and speed with which participants answer questions about text scenarios describing a hypothetical adversary attack, and it was measured subjectively through participant ratings of workload and preference.

A. Participants, Materials, and Procedure

This experiment was identical to the experiment that will be presented in Section III below with the following notable exceptions:

1. This first experiment treated the presence of text markup as a within-subjects manipulation for 100 participants. The second experiment, presented in this paper, treats the presence of text markup as a between-subjects manipulation with 100 participants per condition. This change was made to avoid asymmetric transfer seen between conditions in the first experiment [8].

2. The version of the workload survey used in the first experiment was modified to directly compare both version of

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

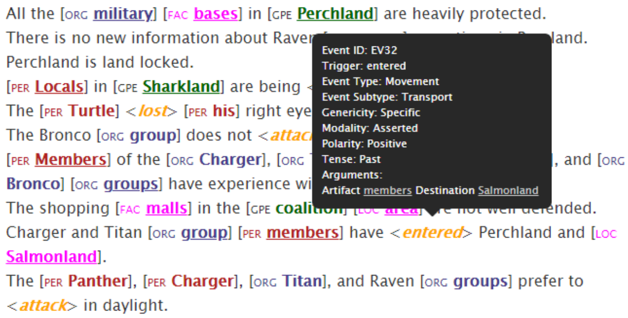

Fig. 1. Excerpt from an ELICIT scenario showing markup with mouse-over information for "entered".

the task (with/without markup). The second experiment, with a between-subject design, used the original workload survey as written.

3. The first experiment did not include a trust in automation survey. The second experiment does.

4. The markup in the first experiment was generated using an IE pipeline developed at Rensselaer Polytechnic Institute [6-7] which highlights a variety of entities (e.g., person, vehicle, geo-political entity) and where mouse-over reveals additional information (e.g., a relation's arguments, the class an entity belongs to). See Fig. 1 for an example of text marked up through this IE pipeline. The markup in the second experiment was created by hand to be as accurate and relevant as possible.

5. The first experiment drew from four separate scenarios. In the second experiment, only two separate scenarios were used. This decision was made for convenience.

For more detail on the materials and procedure for this experiment, see [5] or Section III below.

B. Results

Participants' accuracy and response times are shown for plain and markup trials separately in Fig. 2. While this markup was intended to improve performance, these results point to a small advantage for text without markup over text with markup.

A Wilcoxon signed-rank test determined that participants answered significantly more questions correctly in the Plain condition (median = 5) than in the Markup condition (median = 4.5, p=0.04). Among the 77 participants that showed an asymmetry in their accuracy counts across conditions, 46 (60%) scored higher in the Plain condition.

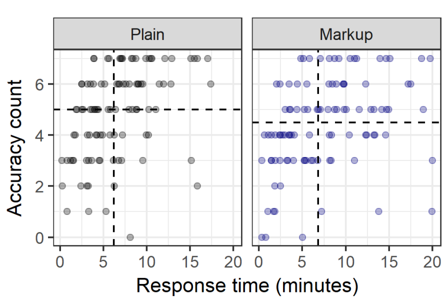

Fig. 2. Accuracy count (number of correctly answered questions) versus response time in minutes for each participant in each condition. Medians are shown as dotted lines.

An additional Wilcoxon signed-rank test determined that participants responded significantly faster in the Plain condition (median = 6.19 minutes) than in the Markup condition (median = 6.83 minutes, p=0.02). Fifty-eight of the 100 participants (58%) responded faster in the Plain condition.

Beyond responding more accurately and quickly on markup trials, participants overall associated higher workload with the markup trials and showed a preference for plain trials. Participant scores on these 21-point scales were binned (1-10, 12-21) to separate responses indicating the plain version of the task from those indicating the markup version of the task (the side of the scale associated with each of these versions was randomized, scores of 11 were dropped from analysis). For each question, Pearson's Chi-squared tests show significantly more favorable scores for the plain version, shown in Table I.

TABLE I. COMPARATIVE WORKLOAD AND PREFERENCE

Question	Number of participants that chose this version of the task	
	Plain	***Markup***
1. *Which version of the task felt more mentally demanding?*	29	64
	$\chi^2(1, N=93)=13.17, p<0.001$	
2. *Which version of the task felt more physically demanding?*	22	45
	$\chi^2(1, N=67)=7.90, p=.005$	
3. *Which version of the task felt more hurried or rushed?*	21	49
	$\chi^2(1, N=70)=11.2, p<0.001$	
4. *On which version of the task do you think you performed better?*	57	34
	$\chi^2(1, N=91)=5.81, p=.016$	
5. *On which version of the task did you feel you had to work harder?*	25	64
	$\chi^2(1, N=89)=17.09, p<0.001$	
6. *Which version of the task lead you to feel more insecure, discouraged, irritated, stressed, or annoyed?*	26	62
	$\chi^2(1, N=88)=14.73, p<0.001$	
7. *Overall, which version of the task you do you prefer?*	66	30
	$\chi^2(1, N=96)=13.5, p<0.001$	

The poorer scores, both objective and subjective, seen on markup trials in this experiment suggest that markup is detrimental to performance. The markup used here, however, is far from perfect and should by no means signal an indictment of all forms of markup for all tasks. To better understand how markup might be of use, the markup in the next experiment was designed to be optimal, i.e., as accurate and as relevant to the task as possible. If participants perform better with such markup than without, further testing can be conducted to estimate thresholds in accuracy and relevance required for markup to be considered a helpful addition, as well as to explore the role or markup density, the number of distinct categories, the way categories are indicated within the text, etc. If participants do not perform better with this markup than without, more serious thought should be given to the use of such markup

III. CURRENT EXPERIMENT

The current experiment tests whether participants show improvements in objective and subjective measures of performance using maximally-accurate and maximally-relevant text markup over non-marked-up text in uncovering a hypothetical adversarial attack.

A. Participants

Two hundred participants were recruited through Amazon Mechanical Turk to take part in this experiment. Each participant was compensated $2.00.

B. Materials

The experiment was prepared using the Ibex tool for running behavioral psycholinguistic experiments (https://code.google.com/archive/p/webspr/) and run online through Amazon Mechanical Turk. The markup used in this experiment was generated by hand by the first author and checked by the other authors, and it separately highlights phrases relevant to four types of answers (*Who*, *What*, *Where*, and *When*) participants are required to provide, as described in the next paragraph. While there are many ways to judge relevance, the researchers believe that highlighting all and only potential answers made the markup much more relevant than the markup in the first experiment without making it too computationally unrealistic or causing it to directly give away any answers. See Fig. 3 for an example of marked-up text. The markup in this experiment drops the bracketing and labeling used in the first experiment, as participants often commented that they found this distracting. Furthermore, the markup here is expressed through background color instead of font color, as we felt this better allowed us to maintain four visually distinct categories (*Who*, *What*, *Where*, and *When*) without sacrificing the contrast between text and background color [3].

The text used in this experiment was drawn from ELICIT, the Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust [9]. ELICIT is a simulated intelligence task containing a number of hypothetical adversary attack scenarios, each in the form of a list of 68 simple sentences that together allow a reader to deduce the *Who*, *What*, *When*, and *Where* of an anticipated adversary attack. These questions are answered in this experiment through seven drop-down menus (*When* is broken down into separate menus for month, date, time of day, and am/pm). See Figs. 1 and 3 for example sentences from ELICIT scenarios.

The Turquoise group focuses on destroying energy infrastructure.

No attacks are being planned on religious organizations in Sigmaland.

The target is in a coalition country (Muland, Xiland, Omicronland, Piland, or Sigmaland).

An attack is being planned for the first month of the year.

The Rose group may be involved.

There is a lot of activity involving the Rose group.

Embassies in Piland were recently attacked and evidence of more attacks has been found.

There are reports that spent nuclear fuel is missing in Muland.

The largest bank in Piland has 4 machine gun emplacements on its roof.

No traces of members from the Blue group have been found in Sigmaland.

Fig. 3. Excerpt from an ELICIT scenario showing hand-generated markup designed to be as accurate and relevant as possible.

This experiment included three questionnaires: a demographic questionnaire, a trust in automation questionnaire, and the NASA Task Load Index (NASA-TLX) [10]. The NASA-TLX asks participants to rate the task on a variety of workload measures, and a question was added asking participants, if they were to do the task again, whether they would prefer to do the task with or without markup. These questions can be seen in Table II. Participants responded to each question by choosing a point on a 21-point scale where 1 represents "very low" or "perfect", and 21 represents "very high" or "failure". The trust in automation questionnaire, proposed by the United States Air Force Research Laboratory [13-15], consisted of 12 questions that capture how the participants feel about automation. Each question is rated from 1-7 on level of agreement (1 representing "disagree" and 7 representing "agree").

C. Procedure

At the beginning of the experiment, participants completed a demographic questionnaire and read a page of instructions explaining the experiment. Participants then completed an abbreviated practice scenario in each condition (Plain and Markup) in order to familiarize them with the scenario presentation and the method for answering questions, as well as to ensure that each participant had some exposure to the markup (important for the trust in automation questionnaire). Participants then completed a trust in automation survey to gauge their baseline trust. At the end of the experiment, participants again completed the trust in automation questionnaire, as well as the workload and preference questionnaire. In addition, feedback was obtained from each participant on the strategies that they used to help them complete the scenarios, and they were given an opportunity to provide comments.

Each participant completed two different test scenarios, both either with markup (Markup condition) or without (Plain condition), determined randomly. Accuracy and response time were collected for each test scenario. Each test scenario was presented along with a countdown timer, and participants were cautioned that their responses would be submitted automatically if 20 minutes elapsed during that scenario.

D. Results

After concerns about quality of responses in the previous experiment, several filtering criteria were established. Unfortunately, this led to a loss of nearly half of all participants. In the results presented below, only the criteria based on response time, requiring participants to have spent at least two minutes and each test scenario, was maintained, as it was the least subjective. This caused 50 participants to be removed from analysis. One additional participant was removed due to a technical failure, leaving 80 participants in the plain condition and 69 participants in the markup condition. All participants' accuracy and responses times are shown in Fig. 4. All other measures include only the filtered 149 participants.

A Wilcoxon rank sum test found no significant difference in the number of correctly answered questions between conditions (Plain median = Markup median = 6, W=10976, p=0.93, r=0.005).

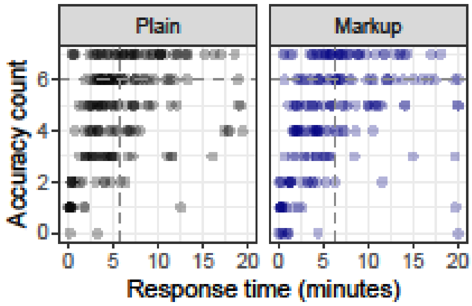

Fig. 4. Accuracy count (number of correctly answered questions) versus response time in minutes for all participants, separated by each condition. Medians (with filtering criterion applied) are shown as dotted lines.

An additional Wilcoxon rank sum test found no significant difference in response time between conditions (Plain median = 5.73, Markup median = 6.38, W=12005, p=0.19, r=0.08). However, while the effect size here is quite small, it suggests that with more power significantly faster response times may emerge in the Plain condition, as was seen in the first experiment.

Responses to the NASA-TLX questions were binned (1-7, 8-14, 15-21) and compared across conditions using Pearson's Chi-squared test. Only the fourth question about Overall Performance emerged as differing significantly between conditions, with responses skewing higher in the markup condition.

TABLE II. WORKLOAD

Question	Number of participants choosing 1-7/8-14/15-21	
	Plain	***Markup***
1. *How mentally demanding was the task?*	0/15/65	2/15/52
	χ^2(2, N=149)=2.65, p=0.266	
2. *How physically demanding was the task?*	67/7/6	56/9/4
	χ^2(2, N=149)=0.82, p=0.662	
3. *How hurried or rushed was the pace of the task?*	29/31/20	20/31/18
	χ^2(2, N=149)=0.95, p=0.621	
4. *How successful were you in accomplishing what you were asked to do?*	35/31/14	25/18/26
	χ^2(2, N=149)=7.95, p=0.019	
5. *How hard did you have to work to accomplish your level of performance?*	1/17/62	2/15/52
	χ^2(2, N=149)=0.53, p=0.769	
6. *How insecure, discouraged, irritated, stressed, and annoyed were you?*	35/24/21	33/25/11
	χ^2(2, N=149)=2.41, p=0.300	

The question on preference, "*If given the choice, which version of the task would you prefer to work with?*", directly compared both versions of the task and so was binned as in the first experiment (1-10, 12-21), where 1-10 = plain preference and 12-21 = markup preference. Responses were pooled across both conditions, and a Pearson's Chi-squared test showed a significant preference for markup (χ^2(1, N=124)=23.52, p<0.001), with 35 participants preferring the Plain condition and 89 participants preferring the Markup condition.

While there are no clear differences among objective measures of performance, the subjective measures hint at higher perceived success with markup and an overall preference for markup. This contrasts with the first experiment, where all advantages were in favor of plain trials.

Table III shows the median responses to the trust in automation questionnaire that was given before and after the test scenarios, where the markup seen in training (and for some participants in test) was described as having been automatically computer generated. Overall, response scores are similar across conditions (Plain vs. Markup). A Pearson's Chi-squared test is used to check for significance between questionnaire responses for the Before and After cases of each condition (Plain, Markup). Results indicate that for both conditions, there is no statistically significant difference between Before and After. While the response scores are all similar, it can be generally seen that the response scores for the positive affinity questions (6-12) decrease after completing the scenarios in the Plain condition, whereas the scores stayed constant or increased after completing the scenarios in the Markup condition. This is a qualitative indication that the Plain condition, which provided the participant with no task-relevant markup guidance, negatively affected how participant view automation (perhaps because they felt markup would not have been helpful on this type of task), and the Markup condition validated the participants' opinions towards automation (perhaps because they did indeed find the markup useful). Across conditions, the low score for the negative questions and the high score for the positive questions suggests that the participants trust automation and would utilize automation to help complete tasks like this. The ability to trust automation for assistance and guidance with tasks allows for increased collaboration performance and reduced cognitive burden, which ultimately leads towards improved situational awareness.

TABLE III. TRUST IN AUTOMATION QUESTIONNAIRE RESULTS

Trust in Automation Questionnaire	Median response			
	Plain		***Markup***	
	Before	***After***	***Before***	***After***
1. *The system is deceptive*	2	2	2	2
2. *The system behaves in an underhanded manner*	2	2	2	2
3. *I am suspicious of the system's intent, action, or outputs*	2	2	2	2
4. *I am way of the system*	3	3	2	3
5. *The system's actions will have a harmful or injurious outcome*	2	2	2	2
6. *I am confident in the system*	5	4	5	5
7. *The system provides security*	4.5	4	5	4
8. *The system has integrity*	5	4	4	5
9. *The system is dependable*	5	4	5	5
10. *The system is reliable*	5	4	5	5
11. *I can trust the system*	5	4	5	5
12. *I am familiar with the system*	5	4.5	5	5

Several strategies were utilized by the participants in each condition to help complete the experimental tasks. Participants in the Markup condition generally used the highlighted text to help them determine the relevant information, used the process of elimination, and focused on easier questions to answer (typically When and Where). While some participants used similar strategies in the Plain condition, it appears that some faced more difficulties and had to adopt different strategies; without markup, helpful strategies included taking notes and using the Find command. While purely qualitative, this suggests that the markup assists and guides participants in an effective way that decreases the workload of completing this task, and without markup, participants will often still seek to use automation (e.g., automated search) to assist them.

IV. Discussion

After participants showed better performance without markup than with markup in a previous experiment, the experiment presented here attempted to create a best-case scenario for markup and show that the current path taken by many IE researchers holds promise for improving SA. The current experiment resulted in no clear preference for text without markup over text with markup, but the only advantage for performance with markup was seen in subjective measures, and with more power an objective advantage for performance without markup may emerge. Still, other qualitative signs emerged that participants prefer to use markup, including a small improvement in trust in automation for participants after using markup and evidence among reported strategies that participants actively used the markup to solve scenarios when it was available to them. While this success for markup is very modest, it suggests further changes that may lead to improved performance and SA, and even if participants only ever show a subjective preference for markup without any improvement in performance, this may be enough to justify its existence.

In this paper, a preliminary look at recently collected data was presented, and there is much further analysis that can be done. Notably, it will be interesting to relate trust in automation scores to preference scores. The researchers expect to find that participants who indicated that they would prefer to do the task in the Plain condition will show lower trust in automation than their counterparts who indicated that they would prefer to do the task in the Markup condition. Additionally, evaluation to explore the role that demographic information (especially occupation) plays in performance and opinion will be conducted in future work. These results may point toward important points of choice and flexibility for automated systems.

There are a range of additional issues that can be addressed in future work and that weigh on the interpretation of the results presented here. Importantly, the markup in this experiment was designed to be as accurate and relevant as possible, but that does not mean it was actually optimal. For example, while not as computationally plausible, all and only correct answers could have been highlighted, providing arguably more task-relevant markup that would have led to better performance in the Markup condition. Additionally, while the researchers believe that our high contrast highlighting was an improvement over the lower contrast font color and noisy bracketing used in the first experiment (cf. Fig. 1 and 3), it cannot be asserted that there is no more advantageous way to present the text. Further steps may include tweaking the text presentation as well as our view of task relevance (both by varying the task and the markup) to explore their impact on performance.

Acknowledgment

Many thanks to Stephen Tratz, Claire Bonial, Jeffrey Micher, Clare Voss, Jon Bakdash, Lucia Donatelli, and Jeff Hoye for their assistance in designing, deploying, and interpreting this work.

References

[1] J. Nielsen, *Designing web usability: The practice of simplicity*, Indianapolis, IN: New Riders Publishing, 1999.

[2] L. Rello, M. Pielot, M. Marcos, “Make it big!: The effect of font size and line spacing on online readability,” Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 3637-3648.

[3] T. R. Williams, “Guidelines for designing and evaluating the display of information on the Web,” Technical Communication, 47(3) 2000, pp. 383-396.

[4] L. Akhmadeeva, I. Tukhvatullin, and B. Veytsman, “Do serifs help in comprehension of printed text? An experiment with Cyrillic readers,” Vision Research, vol 65, 2012, pp. 21-24.

[5] E. Zaroukian, “Information extraction for optimized human understanding and decision making,” Proceeding of the 23rd ICCRTS, Pensacola, FL, November 2018.

[6] Q. Li and H. Ji, “Incremental Joint Extraction of Entity Mentions and Relations,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, June 2014, pp. 402–412.

[7] Q. Li, H. Ji, and L. Huang, “Joint Event Extraction via Structured Prediction with Global Features,” Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, August 2013, pp. 73–82.

[8] E. C. Poulton and P. R. Freeman, “Unwanted asymmetrical transfer effects with balanced experimental designs,” in Psychological Bulletin, 1966, pp. 1-8.

[9] M. Ruddy, “ELICIT – The Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust,” Proceedings of the 12th ICCRTS, Newport, RI, June 2007.

[10] NASA: NASA Task Load Index (TLX), v. 1.0 Manual, 1986.

[11] E. T. Davis and J. Palmer, “Visual search and attention: An overview,” Spatial Vision, vol 17 (4-5), 2004, pp. 249-255.

[12] J. C. Jan, C. M. Chen, and P. H. Huang, “Enhancement of digital reading performance by using a novel web-based collaborative reading annotation system with two quality annotation filtering mechanisms,” International Journal of Human-Computer Studies, vol 86, 2016, pp. 81-93.

[13] A. Neigel, J. Caylor, S. Kase, M. Vanni, J. Hoye, “The Role of Trust and Automation in an Intelligence Analyst Decisional Guidance Paradigm,” Journal of Cognitive Engineering and Decision Making, vol. 12, no. 4, 2018, pp. 239–247.

[14] Jian, J. Y., Bisantz, A. M., and Drury, C. G., “Foundations for an empirically determined scale of trust in automated systems,” International Journal of Cognitive Ergonomics, 4(1), 53-71 (2000).

[15] Kase, S. E., Vanni, M., Caylor, J., & Hoye, J. (2017, May). Human-Assisted Machine Information Exploitation: a crowdsourced investigation of information-based problem solving. In Next-Generation Analyst V (Vol. 10207, p. 1020705). International Society for Optics and Photonics.