Information extraction for optimized human understanding and decision making

Erin Zaroukian
Computational and Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD, USA

ABSTRACT

Information extraction (IE) pipelines aim to point human decision makers toward relevant information, but beyond the accuracy of the pipeline itself, designing the presentation of the output of the pipeline for optimal human understanding should be a goal. This paper establishes a framework for testing comprehension of text documents with and without markup from an IE pipeline and reports the results of a behavioral experiment where an information extraction pipeline, instead of helping, seems to hurt both objective and subjective measures of performance. These results suggest further steps that can be taken toward developing more human-usable IE pipeline outputs.

Introduction

Information extraction (IE) pipelines have been developed as a way to pull relevant information from large document sets. When a decision maker such as a military intelligence analyst has mountains of documents to synthesize in a limited amount of time, a reliable IE pipeline may be an invaluable aid for making selections of relevant documents or portions of documents for further summarization to facilitate decision making.

While the output of an IE pipeline can take many forms, it often provides markup for input texts, identifying, for example, relevant entities and events. Research on these pipelines typically focuses on precision and recall of the computational outputs, and while these are important measures, the end (human) user is often overlooked. What work that has focused on human users typically looks at user experience with highly specific or low-level features such as font size and serifs, e.g., [1-3]. Little research on markup and its effect on reading comprehension is available, and in particular there is a paucity of work comparing text with markup to text without markup to assess what value markup adds. In response, this paper aims to establish a framework for assessing markup by presenting an experiment comparing comprehension of text documents with and without markup from an IE pipeline. Results of such experiments can then lead us toward developing more human-useable IE pipeline outputs.

Starting with simple text documents and the output of an existing IE pipeline, this paper asks: Does markup improve human comprehension of text documents? Comprehension in this experiment 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 (self-reported mental task demands) and preference (preferred document representation with or without markup). Here, marked-up text surprisingly leads to worse comprehension (lower accuracy, slower response times, higher workload ratings, and lower preference ratings) than comparable text without markup. Further work is proposed to gain a clearer understanding of why this pattern emerged and what it means for creating useful markup.

Section I: Methods and Procedure

1) Participants

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

2) Materials and Equipment

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 using an IE pipeline developed at Rensselaer Polytechnic Institute [4][5]. This markup highlights a variety of entities (e.g., person, vehicle, geo-political entity), and 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.

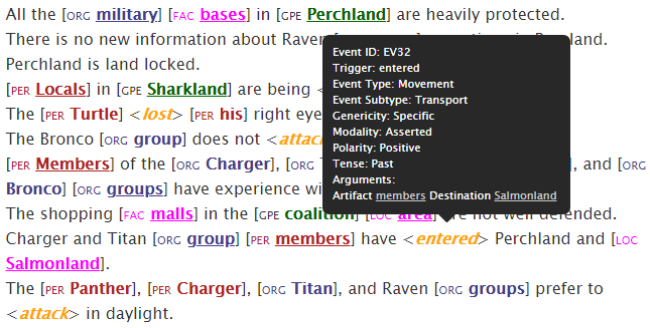

Figure 1: Excerpt from an ELICIT scenario showing markup with mouse-over information for "entered".

The text used in this experiment was drawn from ELICIT, the Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust [6]. ELICIT is a type of hidden profile task, which requires deductive reasoning to solve a problem given different pieces of information. Specifically, 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.[1] These questions are answered in this experiment through seven dropdown menus (*When* is broken down into separate menus for month, date, time of day, and am/pm). See Fig. 1 above for example sentences from an ELICIT scenario.

This experiment included two questionnaires: a demographic questionnaire and a modified version of the NASA Task Load Index (NASA-TLX) [8]. The modified NASA-TLX asked participants to directly compare the two versions of the task (with and without markup) on a variety of workload measures as well as on overall task-version preference. These questions can be seen in Table I. Participants responded to each question by choosing a point on a 21-point scale where the ends of the scale represent a strong preference for each of the versions.

3) Procedure

At the beginning of the experiment, participants completed a demographic questionnaire and read a page of instructions explaining the experiment. Before each test scenario, participants completed an abbreviated practice scenario in order to familiarize them with the scenario presentation and the method for answering questions. At the end of the experiment, participants completed the workload and preference questionnaire. Participants were randomly assigned to see the scale in this questionnaire either with the version with markup on the left and the version without markup on the right, or they were assigned to see the reverse.

[1] See also [7] for work with ELICIT and additional scenarios.

Each participant completed two test scenarios, one with markup (markup condition) and one without (plain condition). The two scenarios were chosen randomly from a set of four scenarios and were assigned randomly to a condition (markup or plain) and order (markup trial first or second). Accuracy and response time were collected for each test scenario.

Section II: Results

Participants' accuracy and response times are shown for plain and markup trials separately in Fig. 2. Overall, these results point to an advantage for text without markup over text with markup.

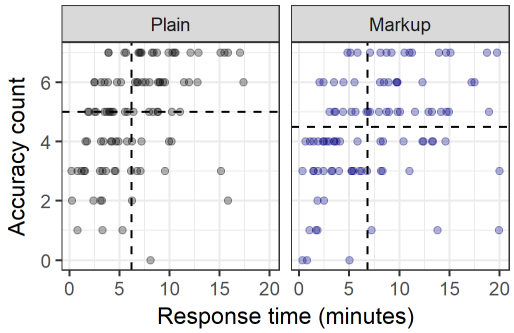

Figure 2: Accuracy count (number of correctly answered questions) versus response time in minutes for each participant in each condition. Medians are shown as dotted lines.

1) Accuracy

Accuracy count (the number of correctly answered questions for a trial, from 0 to 7) are shown on the y axis in Fig. 2 above. A Wilcoxon signed-rank test indicated that participants answered significantly more questions correctly in the plain condition (median = 5) than in the markup condition (median = 4.5, p = 0.04), with 46 out of 77 (60%) participants scoring higher in the plain condition (23 participants scored the same across conditions).

Participants scored similarly in their first (median = 5) and second trials (median = 5), with 45 out of 77 (58%) participants scoring higher on the second trial than on the first. A Wilcoxon signed-rank test indicated no significant difference in accuracy between trials (p = 0.08), showing no clear learning across trials.

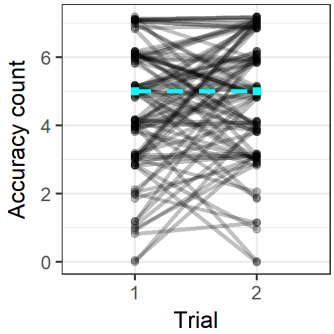

Figure 3: Accuracy count (number of correctly answered questions) for each participant in trial 1 and trial 2. Medians are connected with a dotted line.

The effect of the order of trial condition on participant accuracy is examined in Fig. 4. Among participants whose first trial was in the plain condition, 23 out of 45 (51%) scored higher in the plain condition (median = 5) than in the markup condition (median = 5; 16 participants scored the same across trials[2]). Among participants whose first trial was in the markup condition, 23 out of 32 (72%) scored higher in the plain condition (median = 5) than in the markup condition (median = 4; 7 participants score the same across trials).

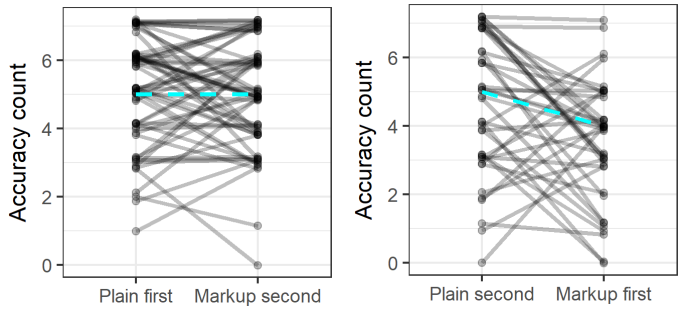

Figure 4: Each participant's accuracy count in plain and markup conditions, connected by a line. Median response times are connected with a dotted line. The plot on the left shows participants who saw the plain trial before the markup trial. The plot on the right shows participants who saw the markup trial before the plain trial.

It is unwise to draw conclusions using only medians from small pools of data, but these results suggest, and further results will support, that a learning advantage is only seen in plain trials: when plain trials are seen second, they seem to benefit from this position, but when markup trails are seen second, they do not experience a similar boost. This asymmetric transfer suggests that, while participants are overall more accurate on plain trials than markup trials, something about them is in some sense hurting later performance.

2) Speed

Response time, or the time elapsed from the beginning of a trial until the participant has selected their answers and clicks the submit button, is shown on the x axis in Fig. 2 above. A Wilcoxon sign-rank test indicated that participants responded significantly faster in the plain condition (median = 6.19 minutes) than in the markup condition (median = 6.83 minutes, $p = 0.02$), with 58 out of 100 (58%) participants responding faster in the plain condition.

As shown in Fig. 5, the median of participants' response times is faster for the second trial (median = 5.93 minutes) than the first trial (median = 6.64 minutes), with 53 out of 100 participants responding faster on the second trial than on the first trial. A Wilcoxon sign-rank test indicated no significant difference in reaction times between trials ($p = 0.52$), again showing no clear sign of learning across trials.

[2] Participants' trial orders were randomly assigned as they launched the experiment, which resulted in 61 participants' first trial being in the plain condition and 39 participants' first trial being in the markup condition. Due to this asymmetry, the overall results above may underrepresent the advantage participants had with in the plain condition. With respect to trial condition order, the data is not only unbalanced, but also non-normally distributed and somewhat sparse, so the decision was made to not run inferential statics within sub-groups.

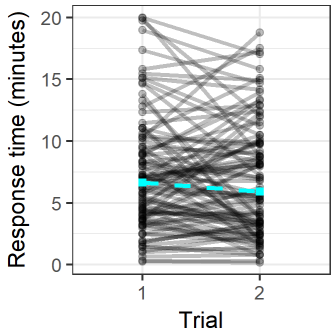

Figure 5: Response time (in minutes) for each participant in trial 1 and trial 2. Medians are connected with a dotted line.

The effect of the order of trial conditions on participant response time is examined in Fig. 6. Among participants whose first trial was in the plain condition, 33 out of 61 (54%) responded faster in the plain condition (median = 6.84 minutes) than in the markup condition (median = 8.05 minutes). Among participants whose first trial was in the markup condition, 25 out of 39 (64%) responded faster in the plain condition (median = 3.90 minutes) than in the markup condition (median = 5.60 minutes).

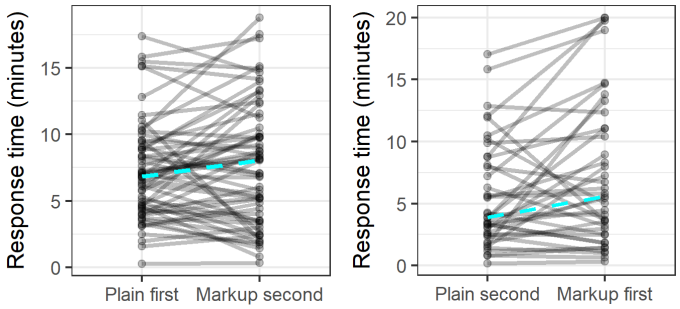

Figure 6: Each participant's reaction time in plain and markup conditions, connected by a line. Median response times are connected with a dotted line. The plot on the left shows participants who saw the plain trial before the markup trial. The plot on the right shows participants who saw the markup trial before the plain trial.

As suggested above, plain trials seem to experience a learning advantage, while markup trails do not.

3) Workload and Preference

Responses to the workload and preference questionnaire are summarized in Table 1, where the Even column gives the percentage of participants who chose the exact middle of the 21-point scale, the Plain column gives the percentage of participants leaning toward the plain version of the task, and the Markup column gives the percentage of participants leaning toward the markup version of the task.

Table 1: Workload and preference.

Question	**Percent of participants that chose this version of the task**		
	Plain	***Markup***	***Even***
Which version of the task felt more mentally demanding?	29	64	7
Which version of the task felt more physically demanding?	22	45	33
Which version of the task felt more hurried or rushed?	21	49	30
On which version of the task do you think you performed better?	57	34	9
On which version of the task did you feel you had to work harder?	25	64	11

Which version of the task lead you to feel more insecure, discouraged, irritated, stressed, or annoyed?	26	62	12
Overall, which version of the task you do you prefer?	66	30	4

This qualitatively shows that participants tended to prefer the plain version of the task and generally associated it with lower workload.

Separating participants into the 66 who prefer the plain version and the 30 who prefer the markup version, shown in Table 2, participants tended to associate a lower workload to their preferred version of the task.

Table 2: Workload and preference by preference.

Question	**Percent of participants that chose this version of the task**					
	Plain preference			**Markup preference**		
	Plain	***Markup***	***Even***	***Plain***	***Markup***	***Even***
Which version of the task felt more mentally demanding?	11	83	6	73	23	3
Which version of the task felt more physically demanding?	8	59	33	57	17	27
Which version of the task felt more hurried or rushed?	11	61	29	47	27	27
On which version of the task do you think you performed better?	80	8	12	7	90	3
On which version of the task did you feel you had to work harder?	8	86	6	67	17	17
Which version of the task lead you to feel more insecure, discouraged, irritated, stressed, or annoyed?	6	86	8	73	7	2
Overall, which version of the task you do you prefer?	100	0	0	0	100	0

Furthermore, descriptively, participants are more accurate on their preferred version of the task, though markup trials are slower for both groups, as summarized in Table 3.

Table 3: Accuracy and speed by preference.

Preference	**N**	**Condition**	**Median accuracy count**	**Median response time (min)**
Plain	66	Plain	6	6.82
		Markup	4	7.15
Markup	30	Plain	4	4.10
		Markup	5	5.25

Looking between these two preference groups, their overall performance differs. Mann-Whitney U tests show that participants who prefer the markup version completed the task significantly faster than participants who preferred the plain version (plain-preference median = 7.06, markup-preference median = 4.48, $p < 0.01$), though their accuracy was not significantly worse (plain-preference median = 5, markup-preference median = 4.5, $p = 0.33$). While this test with relatively small sample sizes has fairly low power, these results reveal some hope for the markup used here, at least with certain participants. Overall, however, participants appear to have found the plain version easier to work with.

Finally, recall that learning transfer, appeared to only occur from markup to plain trials, and not from plain to markup trials. Table 1 shows a tendency to for participants to prefer plain trials, but this tendency is stronger if a participant's first trial is plain: participants whose first trial is plain have a 2.9:1 preference

for plain over markup, whereas participants whose first trial is markup have only a 1.5:1 preference for plain over markup.

Table 4: Percent of participants by trial order and preference.

Trial order	Preference		
	Plain	***Markup***	***Even***
Plain first, markup second	70	25	5
Plain second, markup first	59	38	3

This tendency to develop a preference for the first condition seen, plus an overall preference for plain trials, may cause a resistance to switching conditions between trials that is especially strong when switching from plain to markup. This can be tested in subsequent studies by adding more trials or treating condition as a between subjects manipulation [9]. Additionally, this suggests that the plain preference in Table 1 may give a slightly exaggerated picture, given that 61 out of 100 participants saw their plain trial first; a more accurate picture may come from averaging the columns in Table 4.

Section III: Discussion

This paper presented a framework for evaluating text markup from an IE pipeline, which was then used to evaluate output from an existing IE pipeline. Markup in this experiment, instead of helping, seems overall to hurt performance, both in accuracy and in speed. Additionally, participants tended to find that markup leads to higher workload and is dispreferred in favor of plain, non-marked-up text.

Not all participants performed better without markup, and in fact, descriptively, those who prefer markup are more accurate on markup trials than on plain trials, and while markup trials are slower than plain trials for both groups, participants who prefer markup are overall faster than participants who prefer no markup without suffering from significantly lower accuracy. This human variability points toward the importance of providing the user with means of toggling the markup so that it is only present when it is helpful.

There are a number of additional paths toward better understanding the results presented here as well as how human users can best benefit from the work of an IE pipeline. The markup used in this study was generated automatically by an independently created IE pipeline, so it was not completely accurate, and the categories were not necessarily useful for discovering the *Who*/*What*/*Where*/*When* of the hypothetical adversarial attack. Studies on collaborative annotation suggest that readers perform better with less, higher quality annotation [10]. Similar results were found in [11], where three annotation schemes were compared, and the best performance was found with the simplest, most accurate scheme. In light of this, the next step in this project will be to provide more optimized markup – hand-created markup that is maximally accurate and relevant to the task – to attempt to find better performance with markup than without. If this can be established, it may subsequently be possible to independently vary accuracy and relevance to find thresholds that must be met for markup to be useful. Markup will be a between-subjects between-subjects manipulation in order to avoid asymmetrical transfer between conditions, and additional questions (e.g., occupation, trust in automation) will be added which may shed light on why participants prefer or disprefer markup.

While this experiment used a simple set of texts from which a participant can deduce answers to ELICIT's *Who*, *What*, *Where*, and *When* questions, texts are typically more complicated, do not come with predefined questions, and do not necessarily point to a single answer to any question. It is not obvious that different kinds of texts and tasks would yield the same patterns of results. When considering different ways to mark up these texts, it should be kept in mind that adding markup to text turns reading into a visual search task, and by better aligning markup with what is known about human visual search (e.g., improved use of color and space) [10] and visual analytics, markup may better reflect the hypothesized advantage provided by information extraction. While the current plans for this research framework remain necessarily narrow, many avenues remain in optimizing the output of IE pipelines for human understanding.

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-17-2-0003. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Bibliography

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. 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.
4. 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.
5. 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.
6. M. Ruddy, "ELICIT – The Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust," Proceedings of the 12th ICCRTS, Newport, RI, June 2007.
7. A. Krausman, "Understanding audio communication delay in distributed team interaction: Impact on trust, shared understanding, and workload," Proceedings of the IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Savannah, GA, pp. 1-3.
8. NASA: NASA Task Load Index (TLX), v. 1.0 Manual, 1986.
9. E. C. Poulton and P. R. Freeman, "Unwanted asymmetrical transfer effects with balanced experimental designs," in Psychological Bulletin, 1966, pp. 1-8.
10. E. T. Davis and J. Palmer, "Visual search and attention: An overview," Spatial Vision, vol 17 (4-5), 2004, pp. 249-255.
11. 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.
12. A. Neigel, J. Caylor, S. Kase, M. Vanni, J. Hoye, "The Role of Trust and Automation in an Intelligence Analyst Decisional Guidance Paradigm," Manuscript submitted for publication.