Deductive reasoning with a simulated information extraction pipeline

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

***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 comprehension should be a goal. This paper aims to establish 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.**

Keywords—information extraction, reading, deductive reasoning, visual search, workload, usability

I. Introduction

Information extraction (IE) pipelines have been developed as a way to pull relevant information from large document sets, be they of tweets, news articles, or even videos. 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 and other media for further summarization to facilitate decision making.

While the output of IE pipelines 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. This paper aims to establish a framework for testing comprehension of text documents with and without markup from an IE pipeline. Results of such tests can then lead us toward developing more human-useable IE pipeline outputs.

Starting with simple text documents with ground truth and the output of existing IE pipeline, the paper asks: Does markup improve human comprehension of text documents? Comprehension in this experiment is measured objectively as the speed and accuracy 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 IE). 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.

II. Methods and Procedure

A. Participants

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

B. Materials and Equiptment

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 [1][2]. 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.

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.

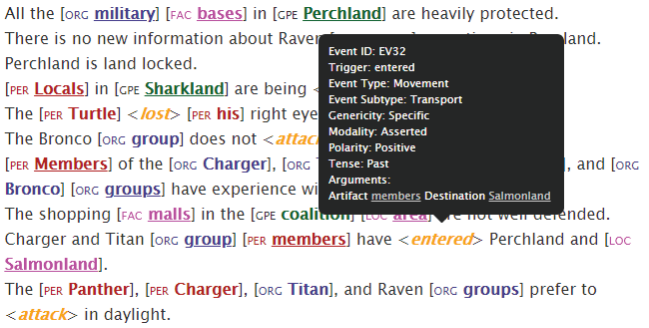

Fig. 1. Exerpt from an ELICIT scenario showing markup with mouse-over information for "entered".

The text used in this experiment was drawn from ELICIT, the Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust [3]. 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 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) [5]. 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 represented a strong preference for each of the versions.

C. 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.

Each participant completed two 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 was collected for each test scenario.

[1] See also [4] for work with ELICIT and additional scenarios.

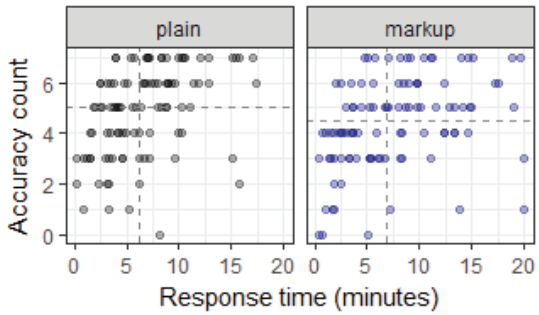

Fig. 2. Accuracy count (number of correctly answered questions) versus response time in minutes for each participant in each condition. Meidans are shown as dotten lines.

III. 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.

A. 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. 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.

The effect of the order of trial condition on participant accuracy is examined in Fig. 3. Among participants whose first

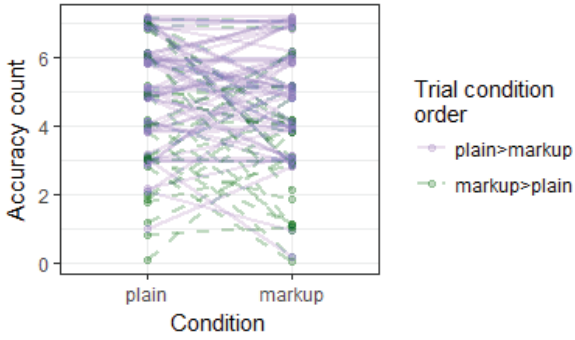

Fig. 3. Each participant's accuracy count in plain and markup conditions, connected by a line, solid and colored purple for participants whose first trial was in the plain consition and dashed and colored green for participants whose first trial was in the markup condition.

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).

When we consider participants who completed their plain trial before their markup trial, we do not see any clear learning transferring from the plain to the markup trial. This may be because any learning is masked by fatigue. When we consider participants who completed their markup trial before their plain trial, we seem to see learning overcoming any fatigue. This suggests that, while participants are overall more accurate on plain trials than markup trials, plain trials may be more fatiguing, leading to poorer accuracy on subsequent trials. This can be tested in subsequent studies by adding more trials or treating condition as a between subjects manipulation [6].

B. *Speed*

Response time is shown on the *x* axis in Fig. 1. A Wilcoxon sign-rank test indicated that participants answered 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.

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 signs of learning across trials.

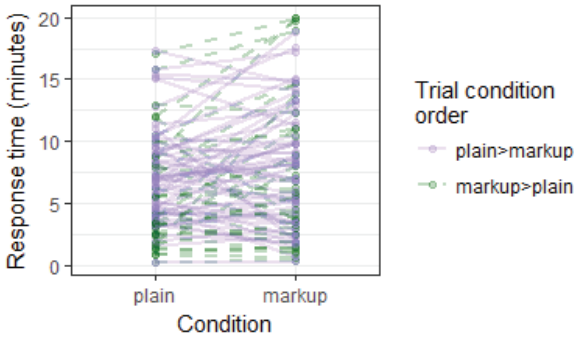

Fig. 4. Each participant's reaction time in plain and markup conditions, connected by a line, solid and colored purple for participants whose first trial was in the plain consition and dashed and colored green for participants whose first trial was in the markup condition.

The effect of the order or trial conditions on participant response time is examined in Fig. 4. 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).

When we consider participants who completed their plain trial before their markup trial, we see slower performance on the second trial. This may be because any learning is masked by fatigue. When we consider participants who completed their markup trial before their plain trial, we now see faster performance on the second trial. This may be because a learning advantage overcomes any fatigue. This suggests that, while participants are overall faster on plain trials, plain trials may be more fatiguing, leading to slower reaction times as well as poorer accuracy on subsequent trials.

C. *Workload and preference*

Responses to the workload and preference questionnaire are summarized in Table I, 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 I. WORKLOAD AND PREFRENCE

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 do you prefer?	66	30	4

While there are clearly differences of opinion between participants, they overall prefer the plain version of the task and associate it with lower workload.

IV. DISCUSSION

Markup in this experiment, instead of helping, seems to hurt performance, both in accuracy and in speed. Additionally,

[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 no inferential statics were run.

participants tend to find that markup leads to higher workload and is dispreferred in favor of plain, non-marked-up text.

While participants perform better without markup, trials without markup may be more taxing in their own way, leading to lower performance on later trials. Not all participants, however, performed better without markup, and in fact we see that, descriptively, those who prefer markup are more accurate on markup trials than on plain trials, though markup always takes longer, as summarized in Table II. This suggests that future comparisons between texts should be made using a between-subjects design to avoid transfer of learning or fatigue.

TABLE II. 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

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. While this experiment used a simple set of texts from which a participant can deduce answers to ELICIT's who, what, when, and where 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 these texts would yield the same pattern of results that we found here. On the other end of the spectrum, a simpler task (e.g., identifying documents containing predefined keywords) could likewise benefit more from markup. Similarly, it is possible that markup may become more helpful in tasks with greater time pressure, after the participant has become more familiar with the markup, or if the IE pipeline achieves higher precision and recall. Finally, 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) [7] and visual analytics, markup may better reflect the hypothesized advantage provided by information extraction.

ACKNOWLEDGMENT

Many thanks are owed to Jennifer Cowley, Jonathan Bakdash, Clare Bonial, Heng Ji, Andrea Krausman, Jeffrey Micher, Shannon Moore, Stephen Tratz, and Claire Voss.

REFERENCES

[1] 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.

[2] 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.

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

[4] A. Krausman, "Understanding audio communication delay in distributed team interaction: Impact on trust, shared understanding, and workload," 2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Savannah, GA, 2017, pp. 1-3.

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

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

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