24th International Command and Control Research & Technology Symposium

Experimentation, Analysis, Assessment, and Metrics

Automated information extraction to facilitate comprehension across text difficulty levels

Erin Zaroukian
US CCDC Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783-1138

Automated information extraction to facilitate comprehension across text difficulty levels

Abstract: Information extraction (IE) pipelines are designed with the end goal of pointing human decision makers toward relevant information within large collections. While research often focuses on the internal computational metrics of the pipeline itself (e.g., F scores), designing the presentation of the output of the pipeline for optimal end user understanding should be a goal. Previous work addressing this goal demonstrated poorer comprehension of text for problem solving when that text was presented with markup from an existing IE pipeline versus plain text, suggesting that this IE pipeline was not well suited for extracting and presenting information to end users. A follow-up study that used markup designed to be maximally accurate and task relevant no longer showed a disadvantage for comprehension with markup, but still failed to show a meaningful advantage. Further investigation into the difficulty of the text showed that text difficultly did not seem to affect performance. Text difficulty, however, is multidimensional and hard to assess, so further testing is required to learn which elements and degrees of text difficulty (e.g., quantity, sparsity of relevant information, degree of dependence between elements) may be relevant and how specific types of markup might facilitate comprehension for different degrees and types of text difficulty.

1. Introduction

1.1. Evaluating automated information extraction

With Information Extraction (IE) research, there is a tendency to focus on extracted information's usefulness to other automated routines downstream, such as ones that take annotated text as input for computing co-reference, translating, populating knowledge bases, or developing watch lists. To evaluate progress in computer science for IE system-building research, the Natural Language Processing (NLP) community typically compares output to datasets curated by humans. IE research has made remarkable progress in this area using this intrinsic-measure framework, relying heavily on precision, recall, and F-score, and most systems for English score very well. In fact, intrinsic metrics are so high for English Named Entity Recognition that many consider IE a solved problem [1].

While incredible progress has been made using these intrinsic metrics for downstream NLP, relatively little attention has been paid to human analysts as downstream processors of extracted information. The project presented in this paper addresses the important issue of what needs to happen to have the technology serve situational awareness, decision making, and other cognitive requirements of human analysts, presenting a framework in which IE systems can be compared against extrinsic metrics. Specifically, participants are presented with either plain text or marked-up text outputs from an IE system that they must use to accomplish a military intelligence task. Their task performance, subjective workload, and preference are compared across text conditions to measure the success of the IE system.

1.2. Evaluating IE with ELICIT

Scenarios from the Experimental Laboratory for Investigating Collaboration, Information-sharing, and Trust (ELICIT) [2] serve as useful text to test the effectiveness of an IE system in military-relevant decision making. ELICIT is a platform for conducting hidden profile experiments, or shared information tasks for decision making. The ELICIT platform controls content distribution and captures shared awareness and mission success, and it comes equipped with four scenarios for distribution among teams. Each scenario is composed of 68 separate sentences, referred to as "factoids" (N.B., each "factoid" is assumed to be true, not a false statement presented as fact), which together allow readers to deduce the perpetrator (*Who*), target (*What*), target country (*Where*), and time (*When*) of a hypothetical planned adversarial attack. Each scenario is built from four types of sentences: Expert, Key, Noise, and Supportive. Expert and Key sentences are needed to solve a scenario, while Supportive sentences help reinforce the solution. Noise sentences do not necessarily provide helpful information. While these scenarios were intended to be distributed across teams, they are also suited for individual decision making, and they are used this way in the experiments presented below.

These scenarios are useful in testing the output of an IE pipeline because they provide a ground truth, allowing easy measures of correctness for participant performance solving a scenario. Further, identifying hypothetical adversary attacks in these scenarios aligns well with military intelligence analyst tasks (though in an actual military context, information is not always reliable, purely deductive reasoning it typically not sufficient, there may be much more noise, etc.), and they are non-sensitive and safe to share with research participants.

1.3. Task difficulty and IE systems

ELICIT scenarios, while similar to each other in structure, vary in difficulty in ways that may have important consequences for assessing the human-usefulness of an IE system. For an easy task, automated assistance may be unnecessary and distracting, while for a very difficult task requiring the analyst to identify and synthesize many disparate pieces of information, reliable clues to what information to focus on may be invaluable in conserving time and working memory and minimizing frustration. As described in [3, p. 44], "Problem difficulty is a reflection of the amount of information and information processing required," but quantifying difficulty is not straightforward.

Among the four included scenarios, or "Factoid sets", the ELICIT software manual claims that "Factoid set 4 is the easiest. Factoid set 1 is next in degree of difficulty. Factoid set 2 and 3 are very similar, but factoid set two has a slight twist that makes it the most difficult factoid set to solve," [2, p. 50] (i.e., 4 < 1 < 3 < 2). No further explanation, however is provided. Work by Anthony Alston [4] as well as Morton & Adams [5] roughly agree with these rankings and provide a variety of difficulty metrics.

1.3.1. Alston

Alston discusses complexity as a component of difficulty, and he considered complexity to be largely determined by the interaction between sentences. Based on two tables, Alston's Table 4, recreated below as Table 1 (with Morton & Adams's results included), and Alston's Table 5 which sums across the four "Factoids per sub-solution", he asserts that "Factoid Set 3 is clearly the most complex, with Table 5 showing Factoid Set 2 nearer in complexity to Factoid Set 3 than Table 4. Both show Factoid Set 4 the least complex by far," (i.e., 4 << 2 < 3, 4 << 1) [4, p. 17].

As shown in Table 1, Alston bases his rankings on four measures: "mixed logic streams" (the number of factoids shared between logic chains required to solve the scenario, see Figure 1, where sentences 2 and 42 are shared between three logic chains), "factoids per sub-solution" (total number of factoids required for each sub-solution, i.e., *Who*, *What*, *Where*, *When*, see Figure 1, where the *What* sub-solution is depicted as requiring 5 sentences)[1], "number of relationships" (I was unable to determine what this means), and "number of factoids" (total number of factoids required to solve the scenario, i.e., all sub-solutions). While these are provided in Table 1, I was unable to replicate these results and was unable to make contact with Alston to gain access to the full analysis. I agree, however, with the general pattern. Alston discusses a number of additional metrics but claims that these are identical across scenarios, though again I disagree in several cases (e.g., he claims that all scenarios have the same number of logical chains/streams per sub-solution, namely, one, cf. Morton & Adams analysis using solution trees below, where all sub-solutions require multiple chains).

Table 1 Table 4 from Alston, including comparable calculations from Morton & Adams.

	Factoid Set 1	**Factoid Set 2**	**Factoid Set 3**	**Factoid Set 4**
Mixed Logic Streams	7	8	9	4
Factoids per sub-solution Who, What, Where, When (Sum)	5, 5, 5, 9 (24)	5, 11, 8, 10 (34)	10, 8, 14, 4 (36)	5, 7, 6, 4 (22)
Factoids per sub-solution – from Morton & Adams solving matrices	*5, 4, 7, 9 (25)*	*5, 5, 4, 9 (23)*	*8, 4, 7, 8 (27)*	*6, 4, 12, 8 (30)*
Factoids per sub-solution – from Morton & Adams solution trees	*5, 5, 8, 9 (27)*	*5, 6, 5, 9 (25)*	*11, 5, 8, 9 (32)*	*6, 6, 11, 9 (32)*
Number of relationships	25	25	27	17
Number of factoids	15	15	16	12
Number of factoids – from Morton & Adams solution trees	*16*	*17*	*16*	*17*

1.3.2. Morton & Adams

Morton & Adams build solution trees (e.g., Figure 1) to understand the structure of the scenarios in order to develop a framework for systematically generating new scenarios. Solution trees show the logical interaction of the sentences in solving a scenario, where the number of branches depicts the non-exclusivity of sentences (one sentence/branch rules out some alternatives but does not determine a unique solution, e.g., Figure 1 shows three branches), and the length of branches depicts conditionality (e.g., the second step in the chain in Figure 1 is only useful with the information provided in the first step, namely that the target must be unprotected). These trees are based on software-generated solution maps but were presumably created by hand.

Morton & Adams, while not directly referencing difficulty, find that "Scenarios 1 and 2 are highly similar, using the same logical sequences, merely changing names of actors, targets, countries, and timings. Scenario 3 is similar to 1 and 2 but with some significant structural differences. The structural differences are primarily demonstrated in a greater dependence on interim conclusions regarding the answers to *What* and *Where*. Scenario 4 is substantially different from the other three scenarios in structure," [5, p. 3]. This characterization differs from the ELICIT manual, which considers 1 easy and 2

[1] See also [2, p. 43].

the most difficult, and groups 2 and 3 together in terms of similarity (without 1). While they do not offer solving trees as any sort of quantitative measure, they offer several qualitative metrics for assessing scenario difficulty, including the interconnectedness of Expert and Key sentences (as shown in the solution trees), as well as features of the Supporting and Noise sentences and how the sentences are presented.

Table 1 above includes two measures of "factoids per sub-solution" from Morton & Adams for comparison. First are the sentence counts they provide from their "solving matrices", which record which Expert and Key sentences are necessary to deduce a solution. A solving matrix for scenario 1's *What* and *Where* is provided in their paper, though it is not clear why sentence 42 was excluded, cf. scenario 1's solving tree. Morton & Adams later suggest, however, that solution trees are a better representation of factoid interactions in arriving at a scenario solution than are solving matrices. Also included in this table are sentence counts from their solution trees, as a solution tree for each scenario was included in their paper, [5, pp. 20-21]. As with the solving matrices, it is not entirely clear how these trees were decided upon and why they differ from the solving matrices in which sentences they consider necessary (again, cf. sentence 42 for scenario 1's *What* and *Where*). Also included in Table 1 are the "number of factoids", or the number of unique factoids Morton & Adams use in each scenario's solution tree.

The solution tree for scenario 1's *What* is recreated from Morton & Adams in Figure 1. This solution requires several undepicted assumptions, including that the target (*What*) is high value, that a preferred target will be chosen, that "in the region" is equivalent to "in coalition member countries", and that if steps are being taken to protect a given target, it can be treated as a protected target. From the provided sentences, a reader can rule out visiting dignitaries (all are protected - Sentence 29) and coalition member embassies (all Tauland and Epsilonland targets are protected - Sentence 22, all Chiland, Psiland, and Omegaland embassies are protected - Sentence 39, and the target is in Tauland, Epsilongland, Chiland, Psiland, or Omegaland - Sentence 42) and conclude that the target is a financial institution.

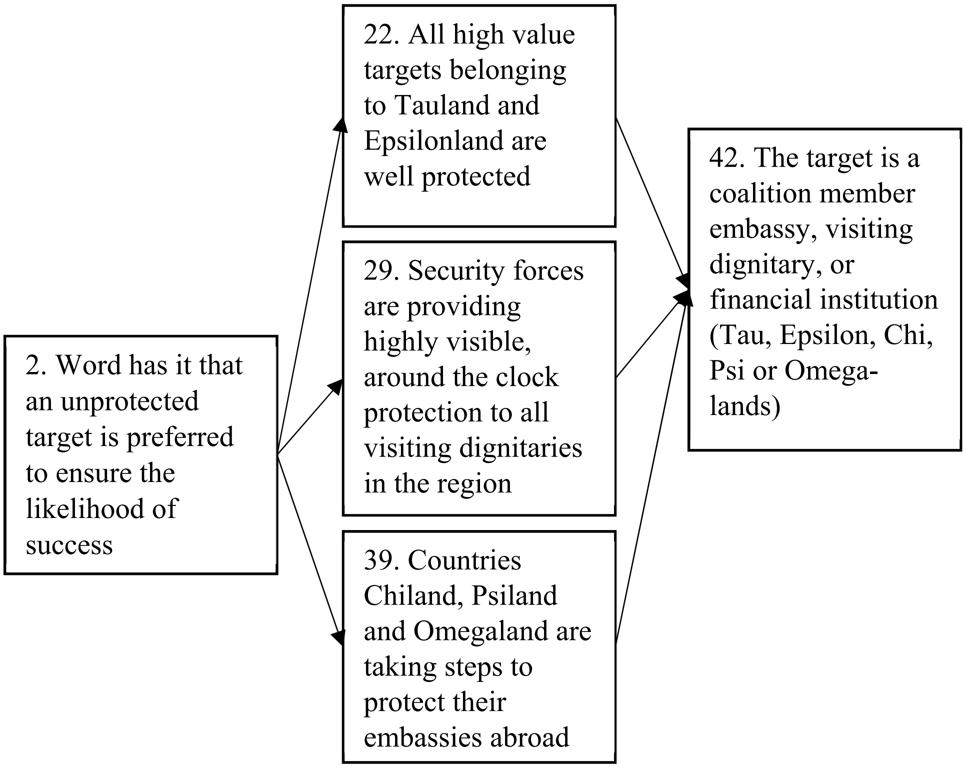

22. All high value targets belonging to Tauland and Epsilonland are well protected
29. Security forces are providing highly visible, around the clock protection to all visiting dignitaries in the region
39. Countries Chiland, Psiland and Omegaland are taking steps to protect their embassies abroad
2. Word has it that an unprotected target is preferred to ensure the likelihood of success
42. The target is a coalition member embassy, visiting dignitary, or financial institution (Tau, Epsilon, Chi, Psi or Omega-lands)

Figure 1 Logic chains for Scenario 1 "What", as given in Morton & Adams.

These two papers demonstrate that there and many dimensions to difficulty and making assessments along any one dimension can be subjective when natural language is being assessed.

2. Experiments

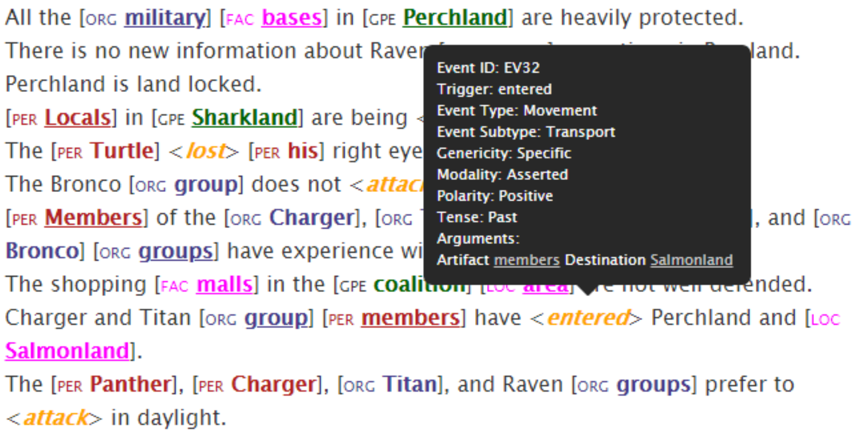
Presented below are two experiments in which human users are presented with full ELICIT scenarios either with or without markup from an IE pipeline and asked to determine the *Who*, *What*, *Where*, and *When* of the hypothetical adversary attack depicted within. Four different Scenarios were used in these experiments: Scenarios 1, 4, 7, and 8. Scenarios 7 and 8 were not included in ELICIT but were created by another researcher by replacing names within Scenarios 3 and 4 respectively, but without changing the structure or, presumably, difficulty. Recall that Scenario 4 (and by extension 8) was generally considered to be the easiest, and among the four scenarios used in these experiments, 7 would be considered the hardest. Scenario 1 was considered to be of intermediate difficulty. This leads to the difficulty ranking 4,8 < 1 < 7.

2.1. Experiment 1

2.1.1. Method

In Experiment 1, described in detail in [6], 100 participants were presented ELICIT scenarios, sets of sentences describing a hypothetical adversarial attack, which they saw plain or with markup from an IE pipeline. The participant's task was to act as analyst and identify the *Who*, *What*, *Where*, and *When* of the attack, and their performance with and without markup from an existing IE pipeline [6-7] was compared to determine whether the markup was helpful.

The markup presented in this experiment shows detected entities (e.g., person, vehicle, geo-political entity) and events (e.g., attack, enter) via bracketing and subscripts, with mouse-over revealing additional information (e.g., an event's arguments, the class an entity belongs to). See Figure 1Figure 2 for an example of ELICIT text marked up through this IE pipeline.



All the [ORG military] [FAC bases] in [GPE Perchland] are heavily protected.
There is no new information about Raven land.
Perchland is land locked.
[PER Locals] in [GPE Sharkland] are being <
The [PER Turtle] <lost> [PER his] right eye
The Bronco [ORG group] does not <attac
[PER Members] of the [ORG Charger], [ORG], and [ORG
Bronco] [ORG groups] have experience wi
The shopping [FAC malls] in the [GPE coalition] [LOC area] are not well defended.
Charger and Titan [ORG group] [PER members] have <entered> Perchland and [LOC
Salmonland].
The [PER Panther], [PER Charger], [ORG Titan], and Raven [ORG groups] prefer to
<attack> in daylight.

Event ID: EV32
Trigger: entered
Event Type: Movement
Event Subtype: Transport
Genericity: Specific
Modality: Asserted
Polarity: Positive
Tense: Past
Arguments:
Artifact members Destination Salmonland

Figure 2 Example markup from Experiment 1.

This experiment also included 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. 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. An additional preference question was included, asking participants which version of the task (with or without markup) they preferred.

At the beginning of the experiment, participants completed a demographic questionnaire and read a page of instructions explaining the experiment. Each participant completed two test scenarios, one with markup (Markup condition) and one without (Plain condition), with order randomized, each preceded by an abbreviated practice scenario. The two scenarios were drawn randomly without replacement from the pool of scenarios 1, 4, 7, and 8. Accuracy and response time were collected for each test scenario. At the end of the experiment, participants completed the workload and preference questionnaire.

2.1.2. Results

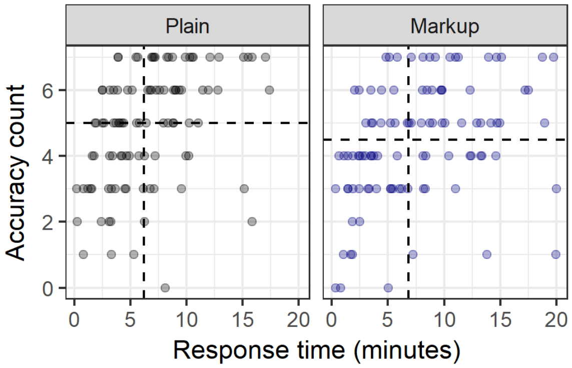


Figure 3 Experiment 1 participant accuracy counts (number of correctly-answered questions, 0-7) and response times for scenarios solved with and without markup. Dashed lines indicate medians.

Participants' accuracy and response times are shown for the plain and markup trials separately in Figure 3. Overall, these results point to an advantage for text without markup over text with markup.

A Wilcoxon signed-rank test indicated that participants answered significantly more questions correctly in the plain condition than in the markup condition, with accuracy counts (the number of correctly identified attack roles for a trial, from 0 to 7) are shown on the y axis in Figure 3. A Wilcoxon sign-rank test indicated that participants completed scenarios significantly faster in the plain condition than in the markup condition, with response time shown on the x axis in Figure 1Figure 3.

Pearson's Chi-squared tests showed that participants overall associated higher workload with the markup trials and showed a preference for plain trials.[2]

Recall that participants completed two scenarios drawn from a set of four: 1, 4, 7, and 8 (4,8 < 1 < 7).

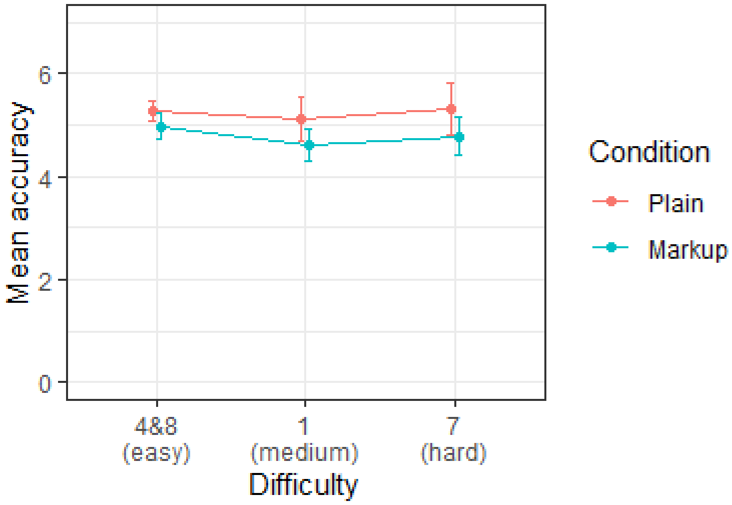


Figure 4 Mean accuracy by Difficulty for Plain and Markup trials in Experiment 1, with standard error.

[2] Again, these results have previously been reported in [6]. These results are confirmed in parametric tests below.

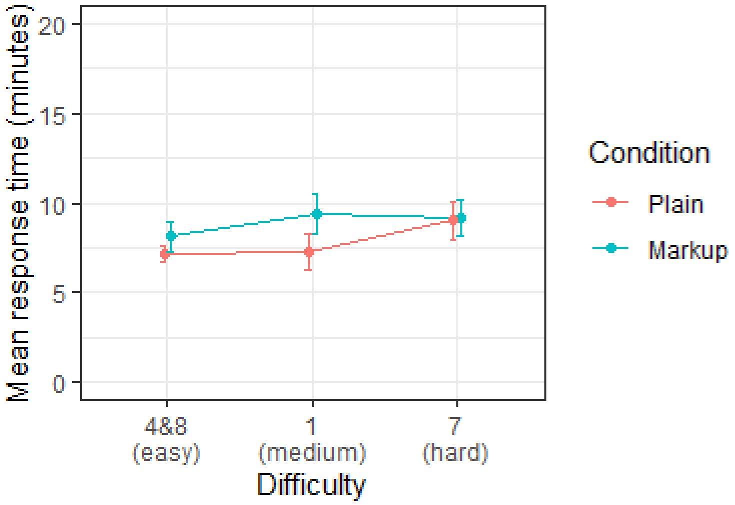


Figure 5 Mean response time by scenario difficulty for Plain and Markup trials in Experiment 1, with standard error.

A linear mixed model predicting response time from condition (Plain, Markup) and scenario difficulty (scenario 4 and 8, scenario 1, scenario 7), with random intercepts for participants was constructed. Difficulty did not affect response time ($\chi^2(1) \ = 0.740, p = 0.691$).

A similar linear mixed model was constructed to predict accuracy from condition and scenario difficulty, with random intercepts for participants, modeling accuracy count as a sequential process [9]. As shown by the Wilcoxon signed-rank test above, condition remains an important predictor (*b* = -0.38; 95% CI = [-0.74, -0.02]; 95% CI excludes zero), but scenario difficulty does not seem to influence accuracy (Scenarios 4 and 8 vs. scenario 1: *b* = -0.14; 95%-CI = [-0.39, 0.11]) (scenarios 4, 8, and 1 vs. scenario 7: *b*=0.06; 95%-CI = [-0.09, 0.22]).

2.1.3. Discussion

While the IE pipeline tested here is intended to help the end user human analyst, this experiment demonstrated the pipeline's markup hurting performance, both in accuracy and speed. Additionally, participants tend to find that markup leads to higher workload and prefer plain, non-marked-up text. It is counterintuitive that markup would be categorically harmful to performance, so there may be forms of markup that are better suited to, and therefore more helpful in, this specific task.

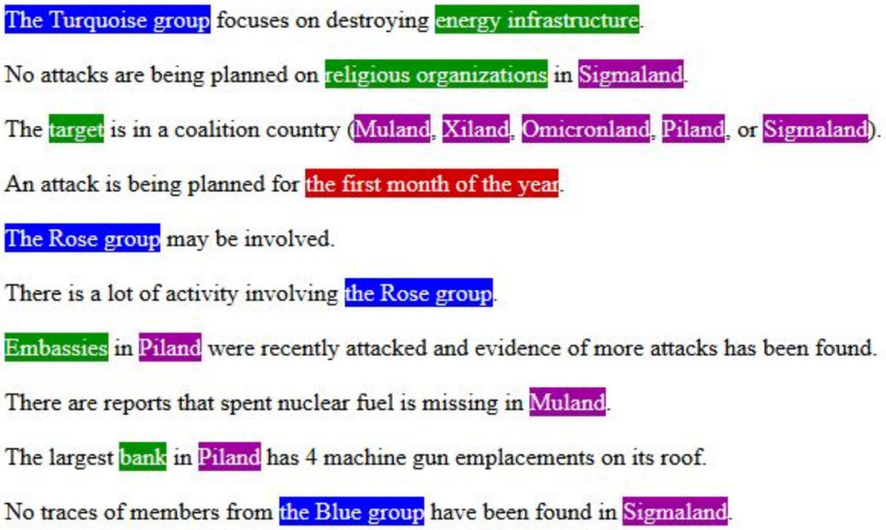
When scenarios where grouped by difficulty, difficulty did not appear to predict response time or accuracy. This may be the difficulty levels considered; while it is not obvious from mean accuracy counts and response times that the difficulty levels have been mis-ranked, the difference in difficulty between levels may be too small to have meaningfully influenced results. Additionally, more information need not improve decision making [10], and the pattern of performance and preference seen in this experiment suggest that participants are experiencing information overload.

2.2. Experiment 2

2.2.1. Method

In Experiment 2, described in detail in [12-12], Experiment 1 was repeated using 100 participants per condition in a between-subject manipulation[3], where each participant either saw two scenarios in the Plain condition or two scenarios in the Markup condition. Scenarios 1 and 4 were used for each participant, with order of presentation randomized.

The markup presented in this experiment was generated by hand such that all and only potential responses to the *Who*, *What*, *Where*, and *When* questions were highlighted. This was done to make the markup more task-relevant than the markup in Experiment 1 (which selected a wider range of entities, very imperfectly) without making it too computationally unrealistic or causing it to directly give away any answers. See Figure 6 for an example of marked-up text.



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.

Figure 6 Example markup from Experiment 2.

This experiment included the NASA-TLX with an additional preference question asking participants which version of the task (with or without markup) they preferred.

At the beginning of the experiment, participants completed a demographic questionnaire and read a page of instructions explaining the experiment. Participants then completed two abbreviated training scenarios, one without markup and one with, to ensure that each participant had exposure to both conditions to allow them to provide an informed preference. Each participant completed two test scenarios, either both with markup or both without. Accuracy and response time were collected for each test scenario. At the end of the experiment, participants completed the workload and preference questionnaire. Additionally, participants in this experiment completed a trust in automation questionnaire, discussed in [12-13].

[3] A between-subjects design was chosen due to asymmetrical transfer seen in Experiment 1.

2.2.2. Results

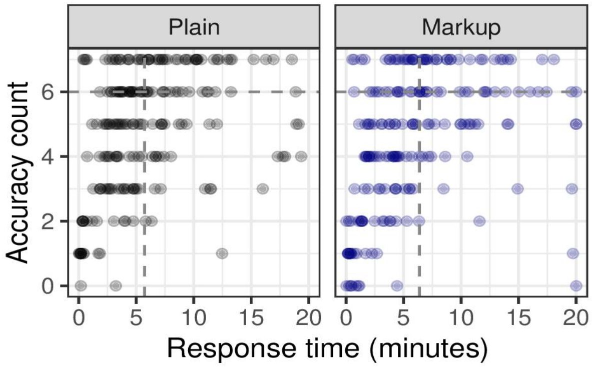


Figure 7 Experiment 2 participant accuracy counts (number of correctly-answered questions, 0-7) and response times for scenarios solved with and without markup. Dashed lines indicate medians.

Participants' accuracy and response times are shown for the plain and markup trials separately in Figure 7. Overall, these results show similar performance for text with and without markup.

Due to a high number of suspected low-quality responses in Experiment 1, it was decided in Experiment 2 that analyses would be run only on participants whose response time on both scenarios was at least two minutes (this amount of time allows for little more than reading the sentences, let along deducing the answers). One additional participant was excluded due to a technical failure. Results reported below therefore include only 149 of the 200 participants (80 Plain, 69 Markup).

A Wilcoxon rank sum test found no significant difference in the number of correctly answered questions between conditions, with accuracy counts shown on the y-axis in Figure 7. An additional Wilcoxon rank sum test found no significant difference in response time between conditions, with response times shown on the x-axis in Figure 7.

Responses to the NASA-TLX differed significantly across conditions using Pearson's Chi-squared test only for the question about Overall Performance ("How successful were you in accomplishing what you were asked to do?"), with responses favoring the markup condition. A Pearson's Chi-squared test showed a significant preference for Markup.[4]

In Experiment 2, participants each saw Scenarios 1 and 4, generally considered the easiest scenarios.

[4] Again, these results have previously been reported in [12-13] and are confirmed in parametric tests below.

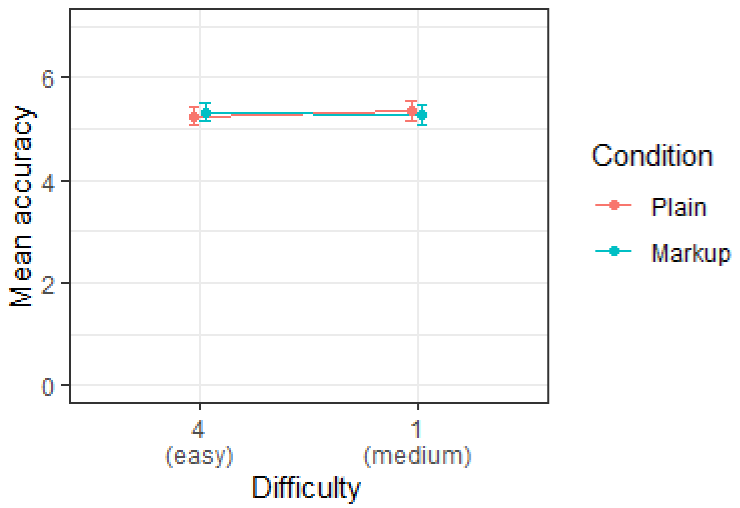


Figure 8 Mean accuracy by scenario difficulty for Plain and Markup trials in Experiment 2, with standard error.

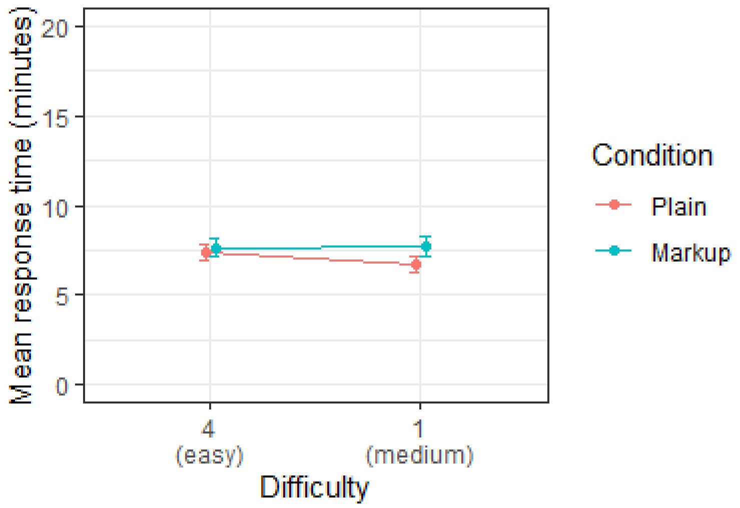


Figure 9 Mean response time by scenario difficulty for Plain and Markup trials in Experiment 2, with standard error.

A linear mixed model predicting response time from condition (Plain, Markup) and scenario difficulty (scenario 1, scenario 7), with random intercepts for participants was constructed. Difficulty did not significantly affect response time ($\chi^2(1) \ = 1.61, p = 0.328$).

A similar mixed model predicting accuracy from condition and scenario difficulty, with random intercepts for participants was constructed, and accuracy count was modeled as a sequential process. Neither condition (*b*=-0.05; 95%-CI = [-0.58, 0.47]) nor scenario difficulty (*b* = -0.13; 95%-CI = [-0.43, 0.17]) seem to influence accuracy.

2.2.3. Discussion

While the markup tested in Experiment 2 did not lead to meaningfully or statistically worse performance, as in Experiment 1, it nonetheless failed to lead to better performance. Similarly, while workload ratings no longer favored the Plain condition, they showed little advantage for the Markup condition. Preference ratings, however, shifted in Experiment 2 to favor the Markup condition.

As in Experiment 1, scenario difficulty did not appear to predict response time or accuracy, though the two scenarios used in Experiment 2 (scenarios 1 and 4) may be of similar difficulty level (cf. Alston [4], who ranks 4 << 1).

3. Conclusion

While the experiments presented in this paper are of limited scope, they demonstrate an experimental framework that can be used to explore further manipulations. These can lead to a better understanding of how various features of tasks and text presentation affect various aspects of performance. Features include task difficulty, which, as seen above, is multidimensional and can be difficult to assess. Additionally, this framework allows various specific use cases to be tested as IE pipelines are developed or as the end user's task changes. This flexibility fosters an IE development loop that includes user testing and user-directed IE development guidelines, promoting systems that succeed not only on intrinsic measures, but on extrinsic measures as well.

While IE as seen in these experiments may be helpful for subtasks, e.g., finding all mentions of the Triangle group or of geo-political entities, it may be less directly helpful for higher-level problem solving. This, in addition to variation seen between individuals' performance and preference, highlight the potential benefits of being about to toggle IE output on and off to match the user's preference and current (sub)-task, while avoiding information overload. Further work may yet find an advantage for marked up text in solving ELICIT scenarios by allowing participants to select the type of markup, including no markup, they wish to use at any point in the task.

References

1. Cunningham, H.: "Information Extraction, automatic." In: Encyclopedia of Language and Linguistics, 2nd Ed., pp. 665-677. Elsevier, New York (2005).
2. Ruddy, M.: ELICIT – The Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust. In: Proceedings of the 12th annual International Command and Control Research and Technology Symposium (2007).
3. Alberts, D.: The Agility Advantage: a survival guide for complex enterprises and endeavors. The Command and Control Research Program Publication, Department of Defense, Washington, DC (2011).
4. Alston, A.: Assessing the Difficulty and Complexity of ELICIT Factoid Sets. In Proceedings of the 15th annual International Command and Control Research and Technology Symposium (2010).
5. Morton, A., and B. D. Adams.: Development of a Team Scenario Content Generation Framework. No. DRDC-TORONTO-CR-2010-105. Humansystems Inc, Guelph, Ontario (2010).
6. Zaroukian, E.: Information extraction for optimized human understanding and decision making. In Proceedings of the 23rd International Command and Control Research and Technology Symposium (2018).
7. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 402–412. ACL, New York (2014).
8. Li, Q., Ji, H., Huang, L.: Joint event extraction via structured prediction with global features. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 73–82. ACL, New York (2013).
9. NASA: NASA Task Load Index (TLX), v. 1.0 Manual. (1986).

10. Bürkner, P. and M. Vuorre. "Ordinal regression models in psychology: a tutorial." *Advances in Methods and Practices in Psychological Science* 2.1 (2019): 77-101.
11. Marusich, L. R., J. Z. Bakdash, E. Onal, M.S. Yu, J. Schaffer, J. O'Donovan, T Höllerer, N. Buchler, C. Gonzalez.: "Effects of information availability on command-and-control decision making: performance, trust, and situation awareness." *Human factors* 58.2 (2016): 301-321.
12. Zaroukian, E., J. Caylor, M. Vanni, and S. Kase: Evaluating Improvement in Situation Awareness and Decision-Making through Automation. In Proceedings of the 24rd International Command and Control Research and Technology Symposium (2019).
13. Zaroukian, E., J. Caylor, M. Vanni, and S. Kase: Human Interacting with the output of information extraction systems. Proceedings of the 10th International Conference on Applied Human Factors and Ergonomics. (2019).