Human Interaction with the Output of Information Extraction Systems

Erin Zaroukian, Justine Caylor, Michelle Vanni, and Sue Kase

Army Research Laboratory, Computational and Information Sciences Directorate,
Adelphi, MD, 20783, USA
{erin.g.zaroukian.civ, justine.p.caylor.ctr, michelle.t.vanni.civ, sue.e.kase.civ}@mail.mil

Abstract. Information Extraction (IE) research has made remarkable progress in Natural Language Processing using intrinsic measures, but little attention has been paid to human analysts as downstream processors. In one experiment, when participants were presented text with or without markup from an IE pipeline, they showed better text comprehension without markup. In a second experiment, the markup was hand-generated to be as relevant and accurate as possible to find conditions under which markup improves performance. This experiment showed no significant difference between performance with and without markup, but a significant majority of participants preferred working with markup to without. Further, preference for markup showed a fairly strong correlation with participants' ratings of their own trust in automation. These results emphasize the importance of testing IE systems with actual users and the importance of trust in automation.

Keywords: Information extraction · Trust in Automation · Reading comprehension · Deductive reasoning · Visual search · Workload · Usability

1 Introduction

With downstream processes for Information Extraction (IE), there is a tendency to consider only automated routines taking annotated text as input for computing co-reference, translating, populating a knowledge base, developing watch lists, or related tasks. Little attention has been paid to human analysts as downstream processors.

To evaluate progress in computer science for IE system-building research, the Natural Language Processing community has compared output to gold-standard datasets curated by humans who have annotated the named entity items in text as being references to entities, in the form of token mentions of IE category types. IE research has made remarkable progress in this area using this intrinsic-measure framework.

Although researchers are always pushing the envelope, most systems for English, trained and tested on standard newswire, do very well, particularly in the area of Named Entity Recognition (NER) within IE systems. Intrinsic metrics are so high for English NER that many consider IE a solved problem [1].[1] This work addresses the important

[1] See [2] though for outstanding issues in NER, such as "different definitions of NE, different types of text, different languages, and noisy data such as OCR and S2T."

issue of what needs to happen to have the technology serve situational awareness, decision making, and other cognitive requirements of human analysts, building a framework in which systems are compared against an extrinsic metric.

2 Experiment 1 - Testing an Existing IE Pipeline

In an experiment described in detail in [3], participants were presented 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 perpetrator, target, time, and location of the attack, and their performance with and without markup was compared to determine whether the markup was helpful.

2.1 Participants

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

2.2 Materials and Equipment

The experiment was created using the Ibex tool for running behavioral psycholinguistic experiments (https://code.google.com/archive/p/webspr/) and run online through Amazon Mechanical Turk.

The text used in this experiment was drawn from the Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust (ELICIT) [4]. ELICIT is a simulated intelligence task containing a number of hypothetical adversary attack scenarios. Each scenario is a list of 68 simple sentences that together allow a reader to deduce the attacker, target, attack time, and attack location (*Who*, *What*, *When*, and *Where*) of an anticipated adversary attack.[2] These roles are identified 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.

The markup presented in this experiment was generated using an IE pipeline developed at Rensselaer Polytechnic Institute [5] [6], which uses NER and event detection techniques. Recognized entities (e.g., person, vehicle, geo-political entity) and events (e.g., attack, enter) were shown via bracketing and subscripts, with mouse-over revealing additional information (e.g., an event's arguments, the class an entity belongs to). See Fig. 1 for an example of ELICIT text marked up through this IE pipeline.

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 as well as overall task-version preference. 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.

---

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

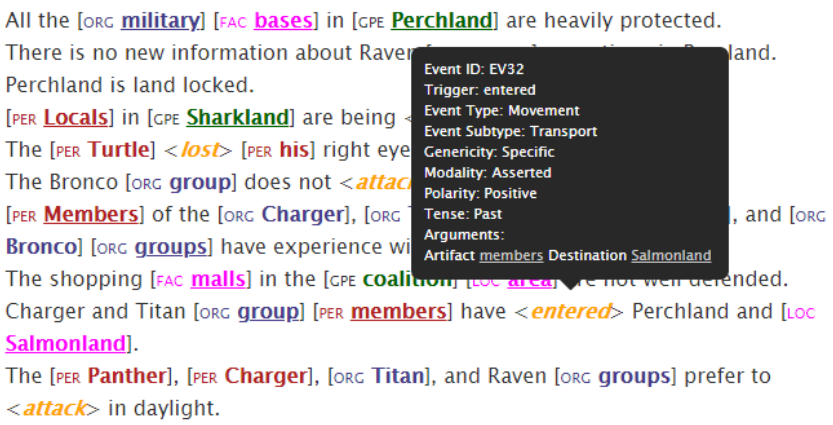

Fig. 1. Excerpt from an ELICIT scenario showing markup with mouse-over information for "entered".

2.3 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 to familiarize them with the task.

Each participant completed two test scenarios, one with markup (Markup condition) and one without (Plain condition), each preceded by an abbreviated practice scenario. Accuracy and response time were collected for each test scenario. At the end of the experiment, participants completed the workload and preference questionnaire.

2.4 Results

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

Accuracy counts (the number of correctly identified attack roles 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).

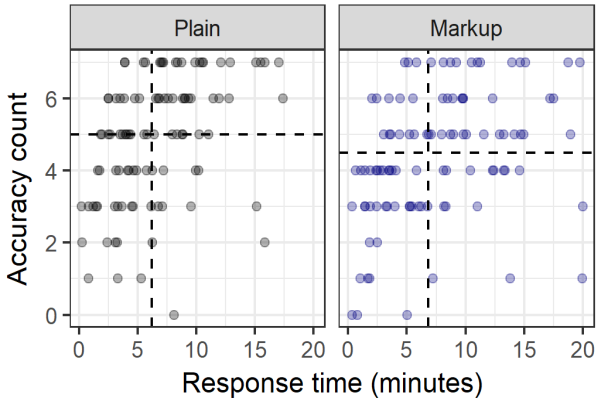

Fig. 2. Accuracy count (number of correctly identified attack roles) versus response time in minutes for each participant in each condition. Medians are shown as dotted lines.

Response time is shown on the x axis in Fig. 2. A Wilcoxon sign-rank test indicated that participants completed scenarios 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.

Workload and Preference. In the interest of space, workload scores will not be discussed here, but they are consistent with preference responses, see [3] for details. The question on preference, "*Overall, which version of the task do you prefer?*", directly compared both versions of the task and so was binned as Plain preference versus Markup preference (scores of 11, indicating no preference, were excluded from analysis). A Pearson's Chi-squared test showed a significant preference for plain over markup ($\chi 2(1, N=96)=13.5, p<0.001$), with 66 participants preferring the Plain condition and 30 participants preferring the Markup condition.

While there is an overall preference for Plain trials, there is still a sizeable minority who prefer Markup trials, and, descriptively, participants prefer the version of the task that they performed better at, as summarized in Table 1.

Table 1. 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

Furthermore, Mann-Whitney U tests shows that participants who prefer the Markup version completed the task significantly faster than participants who preferred the Plain version ($p < 0.01$), though their accuracy was not significantly worse ($p = 0.33$). While

this test with its relatively small sample size 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.

2.5 Discussion

While the IE pipeline tested here is intended to help the downstream human analyst, in this experiment, the pipeline's markup seems to hurt performance, both in accuracy and speed. Additionally, participants tend to find that markup leads to higher workload and is dispreferred in favor of 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.

Additionally, not all participants preferred and performed better without markup. This points toward the importance of providing options to participants, and it may be valuable to identify predictors for whether participants will work well with markup.

3 Experiment 2 - Testing an Ideal IE Pipeline

For this experiment, the aim is to design more relevant and accurate markup for ELICIT scenarios in an attempt to find conditions under which markup improves performance. Further, additional questions are included to provide predictive insight in determining which participants would prefer and perform better with or without markup.

3.1 Participants

This experiment treated Plain/Markup as a between-participants manipulation, so 200 participants were recruited through Amazon Mechanical Turk. Each participant was compensated $2.00.

3.2 Materials and Equipment

Like Experiment 1, this experiment was created using the Ibex tool for running behavioral psycholinguistic experiments (https://code.google.com/archive/p/webspr/) and run online through Amazon Mechanical Turk.

The text used in this experiment is the same text drawn from ELICIT used in Experiment 1.

The markup used in this experiment was generated by hand by the first author and checked by the other authors. It highlights phrases relevant to four types of responses (*Who*, *What*, *Where*, and *When*) that participants are required to provide. While there are many ways to judge relevance, the decision was made to highlight all and only potential responses (e.g., all and only country names were highlighted as possible *Where*s). This strategy was chosen to make the markup 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 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 [9], and the bracketing and labeling used in the first experiment were dropped as participants often commented that they found this distracting.

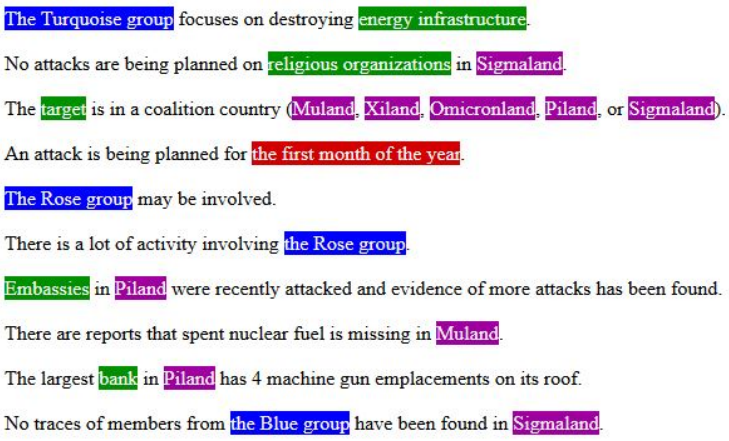

Fig. 3. Excerpt from an ELICIT scenario showing hand-generated markup designed to be as accurate and relevant as possible.

Like Experiment 1, this experiment included a demographic questionnaire, but an additional question about participant occupation was included in hope of finding correlations between reported occupation and preference. Additionally, participants were required to enter a free-text response at the end of the experiment describing any strategies they used to solve the scenarios. This experiment also included an unmodified version of the NASA-TLX (because Plain/Markup was a between-participants manipulation, it would be difficult for participants to directly compare both conditions, so they only rated the version of the task that they completed). A preference question was again included, asking participants whether, were they to participate again, they would prefer the text to appear plain or with markup, indicating their preference by choosing a point on a 21-point scale where the ends of the scale represent a strong preference for each of the versions.

3.3 Procedure

As in Experiment 1, participants first completed a demographic questionnaire and read a page of instructions explaining the experiment. These instructions specified that any markup they see in the experiment was automatically generated (though in this experiment it was actually generated by hand). Participants then completed two practice scenarios, first in the Plain condition, then in the Markup condition. They then completed a Trust in Automation survey [10] asking for subjective ratings about systems like the one that generated the markup seen in the Markup practice scenario. Each participant

completed two test scenarios, both in either the Markup or Plain condition. Accuracy and response time were collected for each test scenario. At the end of the experiment, participants again completed the Trust in Automation survey, provided their strategy descriptions, and completed the workload and preference questionnaire.

3.4 Results

Accuracy and Response Time. Participants' accuracy and response times are shown for plain and markup trials separately in Fig. 4. While Experiment 1 showed an advantage for text without markup over text with markup, the differences here are minimal.

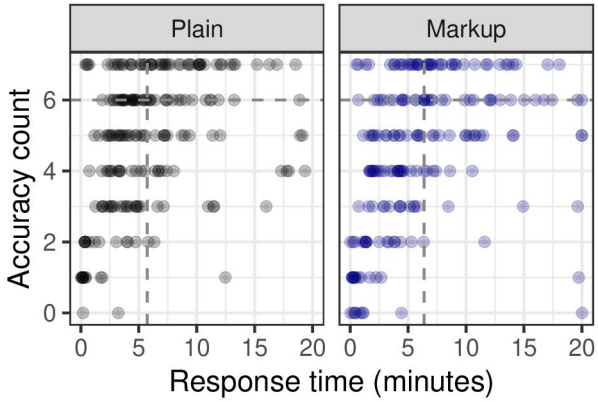

Fig. 4. Accuracy count (number of correctly answered questions) versus response time in minutes for all participants, separated by each condition. Medians (with the filtering criterion applied) are shown as dotted lines.

Concerns about speed and consequent quality of responses were raised in Experiment 1, so for the analyses below only participants with response times of 2 minutes or longer for each test scenario were included (150 out of 200 participants). One additional participant was removed due to a technical failure, leaving 80 participants in the Plain condition and 69 participants in the Markup condition.

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

An additional Wilcoxon rank sum test found no significant difference in response time between conditions (Plain median = 5.73, Markup median = 6.38, $W=12005$, $p=0.19$, $r=0.08$). 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 Experiment 1.

Workload and Preference. Again, in the interest of space, workload scores will not be discussed here, be they were overall similar across conditions. The question of preference, "*If given the choice, which version of the task would you prefer to work with?*", was again binned as Plain preference versus Markup preference (with scores of 11 excluded from analysis). Responses were pooled across both conditions, and a Pearson's Chi-squared test showed a significant preference for Markup ($\chi 2(1, N=124)=23.52$, $p<0.001$), with 35 participants preferring the Plain condition and 89 participants preferring the Markup condition. This contrasts with the first experiment, where all advantages were in favor of Plain trials.

Demographics and Correlations. Participant responses to questions about whether their native language is English, their gender, their age, their level of education, and their occupation are summarized in Table 2.

Table 2. Summary of responses to demographic questions.

Question	Response category	N (All)	N (Filtered)
Language	No	27	16
	Yes	172	133
Gender	Female	86	66
	Male	112	82
	Other/prefer not to say	1	0
Age	18-29	78	45
	30-49	91	78
	50-64	25	22
	65+	5	4
Education	High school graduate or less	20	15
	Some college	70	55
	College degree or more	109	79
Occupation	Science and technology	67	56
	Arts, entertainment, and media	9	7
	Education	7	5
	Legal	13	13
	Sales	17	10
	Food preparation and serving	12	6
	Office administration and support	3	3
	Accounting and finance	20	17
	Healthcare and medical	7	2
	Industry and manufacturing	18	11
	Law enforcement	15	9
	Business management	11	10
	Other	0	0

Neither performance nor preference correlated well with any of the demographic information collected, with the exception of Language, where there is a medium positive correlation between accuracy and being a native speaker of English, as well as a medium negative correlative between response time and being a native speaker of English. Correlation coefficients are shown in Table 3, where one participant (Other/prefer not

to say) was dropped from Gender correlations, and Occupation was pooled into Non-science and technology versus Science and technology.

Table 3. Correlation coefficients between collected demographic information and performance and preference for Plain and Markup trials. A coefficient is listed as 0 if it is less than 0.01 and greater than -0.01.

Covariates		Plain	Markup
Language	Accuracy	0.31	0
	Response time	-0.31	-0.31
	Preference	-0.08	-0.06
Gender	Accuracy	-0.10	-0.10
	Response time	0	-0.14
	Preference	0.05	0.01
Age	Accuracy	0.05	0.03
	Response time	0.03	0.14
	Preference	-0.22	-0.13
Education	Accuracy	-0.05	-0.20
	Response time	0.06	0.10
	Preference	-0.05	-0.18
Occupation	Accuracy	0.12	0
	Response time	0.21	0.06
	Preference	0.04	0.01

Preference for markup showed a fairly strong correlation with participants' ratings of their own trust in automation ($r = 0.39$). The correlation between trust in automation and objective performance measures, however, is very small (accuracy: $r = 0.06$, response time: $r = -0.05$).

3.5 Discussion

Experiment 1 asked participants to uncover hypothetical adversary attacks described in text documents with and without markup from an existing IE pipeline and found that, instead of helping, markup hurt performance and was dispreferred to plain text. While the markup used in Experiment 2 was hand-generated to be as helpful but realistic as possible, it still did not lead to better performance than plain text. This is an important warning to researchers trusting that actual automated markup will be helpful. This markup, however, was overall preferred to plain text, which is valuable for the overall user experience.

These results also emphasize that the trust in the automation that is used in an IE pipeline may be important for user experience and for encouraging users to opt to use these pipelines. However, the link between trust in automation and objective performance measures in the current study is very small, and experiments like this demonstrate that the automation need not improve performance. Much remains to be understood about the gap between IE technology and its human user for this technology to truly support human-computer interaction.

An additional consideration was highlighted by the unexpectedly high number of low-quality responses. These were responses that were too quick to represent true attempts to read the texts and identify the hypothetical adversary attack. The participants providing these responses were roughly twice as likely to report that English was not their native language (11/27 Non-native English speakers were filtered versus 39/172 native English speakers, shown in Table 3), and they often provided incoherent free-text strategy descriptions. This might indicate that workers on Mechanical Turk are generally not willing to put in the work necessary to do well at this task. However, regardless of their performance, this population of workers does not necessarily predict the performance of any other population, importantly, intelligence analysts. While workers on Mechanical Turk can be helpful due to their availability, it is important to include the specific intended end user in the testing cycle.

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

References

1. Cunningham, H.: "Information Extraction, automatic." In: Encyclopedia of Language and Linguistics, 2nd Ed., pp. 665-677. Elsevier, New York (2005)
2. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named entity recognition: fallacies, challenges and opportunities. Computer Standards and Interfaces, 35, pp. 482-489 (2013)
3. Zaroukian, E.: Information extraction for optimized human understanding and decision making. In: Proceedings of ICCRTS (2018)
4. Ruddy, M.: ELICIT – The Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust. In: Proceedings of the 12th ICCRTS (2007)
5. 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)
6. 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)
7. Krausman, A.: Understanding audio communication delay in distributed team interaction: Impact on trust, shared understanding, and workload. In: Proceedings of the IEEE CogSIMA Conference, pp. 1-3. IEEExplore (2017)
8. NASA: NASA Task Load Index (TLX), v. 1.0 Manual. (1986)
9. Williams, T. R.: Guidelines for designing and evaluating the display of information on the Web. Technical Communication, 47(3), 383-396 (2000)
10. 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)