Human interaction with the output of IE systems.

When considering downstream processes for Information Extraction (IE), there is the tendency to consider only automated routines that take annotated text as input for computing co-reference, translating or transliterating, populating a knowledge base, developing watch lists, etc. IE research has made remarkable progress in Natural Language Processing using intrinsic measures, comparing output to gold standard, human curated datasets. Most systems for English, trained/tested on standard newswire, do very well, with precision/recall/F-measure high enough that many consider English IE a solved problem. Little attention, however, has been paid to human analysts as downstream processors.

The work presented below addresses the important issue of what needs to happen to have this technology serve situational awareness, decision-making, and other cognitive requirements of human analysts. This amounts to a framework in which systems are compared against an extrinsic metric – human performance and preference. This will help to bridge an important, and heretofore largely ignored "habitability" gap in the development of IE technology that supports human-computer interaction and human teaming with automata, an exciting and valuable contribution to the Army and beyond.

In a previous study, participants were presented sets of sentences describing a hypothetical adversarial attack [Ruddy 2007], which they saw plain or with markup from an IE pipeline [Li and Ji 2014; Li, Ji, and Huang, 2013]. The participant’s task was to act as analyst and identify the perpetrator, target, time, and location of the attack. Overall, participants showed worse performance with markup than without. Most participants reported preferring the plain text to the markup, but a sizeable minority subjectively preferred and objectively performed more accurately with markup than without.

For the current study, we aimed to design more relevant and accurate markup in an attempt to find conditions under which markup improves performance. Further, we included additional questions that we hoped would be predictive in determining which participants would prefer and perform better with markup or without.

While the current study failed to show better performance with markup than without, it no longer showed a significant pattern in the opposite direction. A significant majority of participants preferred working with markup to without, though, participants tended to prefer whichever condition that they performed well under. Neither preference nor performance correlated well with any of the demographic information collected (gender, age, education, occupation), but preference for markup showed a fairly strong correlation with participants ratings of their own trust in automation (r = .39) [Jian, Bisantz, and Drury 2000].

These results emphasize that 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 (accuracy: r = 0.06, response time: r = -0.05), and experiments like this demonstrate that the automation need not improve performance. Much remains to be understood in the gap between IE technology and its human user in order for this technology to truly support human-computer interaction.

1. Correlation between preference and performance, but is not so strong
   1. If you did well, you tend to prefer the version you did well with. If you did poorly, you tend to prefer the other version.
   2. (Accuracy plain: -.24, markup: .23; RT plain -.19, markup: .16). <- I think these are non-filtered
   3. Could do same/switch – acc .18 / .29, rt .17 / .15
   4. So it’s important to know what you care about
2. Preference/perf ~ occupation = low, as with other demos
   1. Filtered performance: Acc: .07, RT: .13
   2. Filtered preference: <.01 (plain: .04, markup: .01) Tiniest positive correlation between Markup preference and “Science and Technology”
3. Preference/perf ~ trust in automation pretest higher
   1. Not surprising
   2. Filtered performance: Accuracy: 0.06, RT: -0.05 <- not much here
   3. Filtered preference: 0.39 <- pretty big