



U.S. ARMY RESEARCH, DEVELOPMENT AND ENGINEERING COMMAND

Information extraction for optimized human understanding and
decision making

Erin Zaroukian

Computational and Information Sciences Directorate

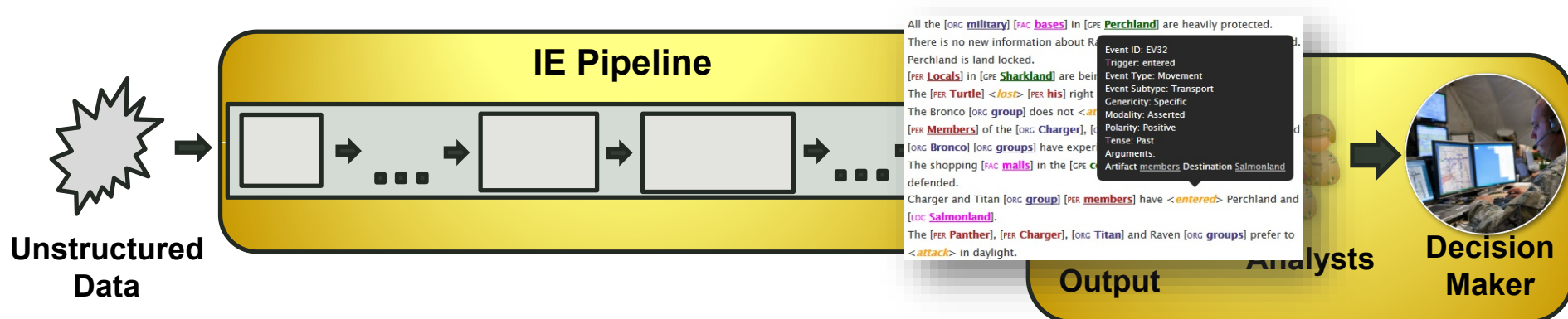
US Army Research Laboratory



U.S. ARMY
RDECOM
ARL RESEARCH LABORATORY

INTRODUCTION

- Information Extraction (IE) pipelines can aid decision making by structuring data and pulling decision-relevant information from large document sets
- Much research focuses on precision/recall of the pipeline
- Little research on how useful the pipeline output is to a user
 - Does text markup from an IE pipeline improve human comprehension of text documents?





RESEARCH QUESTION

Does text markup from an IE pipeline improve human comprehension of text documents?



METHOD

Does **text** markup from an IE pipeline improve human comprehension of **text documents**?

- ELICIT: Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust [Ruddy 2007]
- Scenario
 - 68 sentences
 - Together provide *who/what/where/when* of an anticipated adversary attack

Example mini-scenario

The Tetrahedron group is not involved.

There will be a suicide bomb attack at a new hotel.

The Cube and Tetrahedron groups have the capacity to operate in Gibbonland and Lemurland.

The Cube group only attacks on New Year's Day.

A hotel was recently built in Gibbonland.

The attack will be at 11:00 pm.

Time remaining: 19:31

Terrorist plot:

Who?

What?

Where?

When? at

[→ Click here to submit answers for this scenario](#)



METHOD

Does text markup from an IE pipeline improve human comprehension of text documents?

- Markup generated by an RPI pipeline [Li, Ji, and Huang 2013; Li and Ji 2014]
 - Events, labeled entities, mouse-over

Plain scenario excerpt

All the military bases in Perchland are heavily protected.
 There is no new information about Raven group operations in Bassland.
 Perchland is land locked.
 Locals in Sharkland are being recruited.
 The Turtle lost his right eye in an accident.
 The Bronco group does not attack on its Sabbath.
 Members of the Charger, Titan, Steeler, Raider, and Bronco groups have experience with chemical weapons.
 The shopping malls in the coalition area are not well defended.
 Charger and Titan group members have entered Perchland and Salmonland.
 The Panther, Charger, Titan, and Raven groups prefer to attack in daylight.

Markup scenario excerpt

All the [ORG military] [FAC bases] in [CPE Perchland] are heavily protected.
 There is no new information about Raven group operations in Bassland.
 Perchland is land locked.
 [PER Locals] in [CPE Sharkland] are being recruited.
 The [PER Turtle] <lost> [PER his] right eye in an accident.
 The Bronco [ORG group] does not attack on its Sabbath.
 [PER Members] of the [ORG Charger], [ORG Titan], [ORG Steeler], [ORG Raider], and [ORG Bronco] [ORG groups] have experience with chemical weapons.
 The shopping [FAC malls] in the [CPE coalition 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



U.S. ARMY
RDECOM
ARL RESEARCH LABORATORY

METHOD

Does text markup from an IE pipeline improve human comprehension of text documents?

- Measured objectively as the accuracy and speed with which participants answer questions about the text
- Measured subjectively through ratings of workload and preference

[Type Mention]: Entity
[Type Mention]: Entity with relation
<Trigger>: Event
[PER]: Person [ORG]: Organization [CPE]: Geo-political Entity [FAC]: Facility
[VEH]: Vehicle [WEA]: Weapon [LOC]: Location [OO]: Others
[TIMEX]: Timex [VALUE]: Value

The [per Eagle] is involved.
The Eagle does not work in [cpe Spiderland].
The northern cruise [fac terminal] is ocean-based.
The <attack> will be at the end of the second shift.
The [per Circle] [per group] is recruiting [per locals] - intentions unknown.
The largest [orc museum] in [cpe Spiderland] has a flat roof.
The [per attackers] are focusing on a high visibility target.
The [orc Oval] and [per Hexagon] [orc groups] want to <attack> the interests of [per Hornetland].
The [per Hornetland] will <attack> in [cpe Spiderland].

Time remaining: 19:54

Terrorist plot:
Who?
What?
Where?
When? at

[-- Click here to submit your answer and continue](#)

	Definitely with markup	Equal	Definitely without markup
	-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10		
Which version of the task felt more mentally demanding?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which version of the task felt more physically demanding?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which version of the task felt more hurried or rushed?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On which version of the task do you think you performed better?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
On which version of the task did you feel you had to work harder?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which version of the task lead you to feel more insecure, discouraged, irritated, stressed, or annoyed?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, which version of the task do you prefer?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[-- Click here to continue](#)



PARTICIPANTS AND PROCEDURE

– Participants

- 100 Turkers
- \$2

– Procedure

- Demographic questionnaire
- Instructions
- Plain: Practice scenario, test scenario, answers
- Markup: Practice scenario, test scenario, answers
- Workload and preference questionnaire

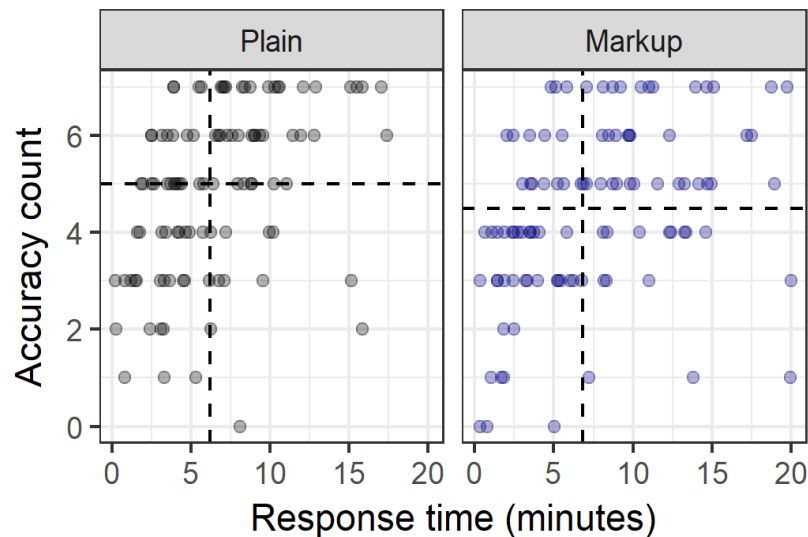
Condition order
randomized





U.S. ARMY
RDECOM
ARL RESEARCH LABORATORY

RESULTS



– Accuracy

- **Plain** $>^*$ **Markup** (med = 5, 4.5)
 - 60% of participants (46/77)

– Speed

- **Plain** $<^*$ **Markup** (med = 6.19min, 6.83min)
 - 58% of participants (58/100)

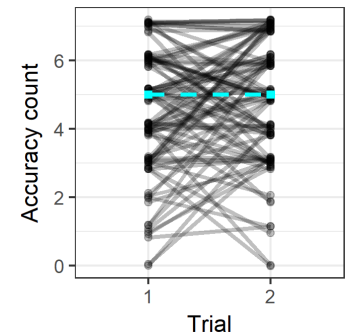
* Significant by Wilcoxon signed-rank test



RESULTS - ACCURACY

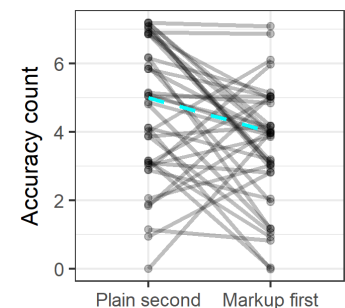
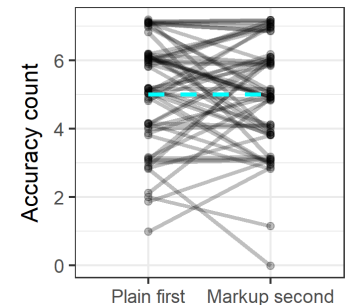
– Trial order

- Second trial $>_{ns}$ First trial (med = 5, 5)
 - 58% of participants (44/77)
 - *No clear learning between first and second trials*



– Condition order

- Plain first: Plain \geq Markup (med = 5, 5)
 - 51% of participants (23/45)
- Plain second: Plain \geq Markup (med = 5, 4)
 - 72% of participants (23/32)
 - *Learning from Markup to Plain?*
(Asymmetric transfer)

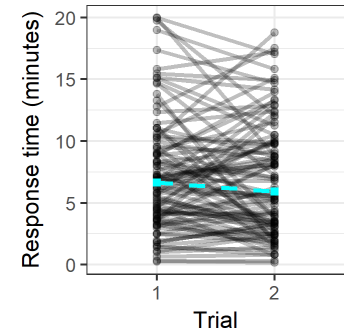




RESULTS - SPEED

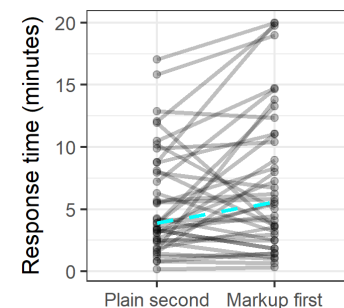
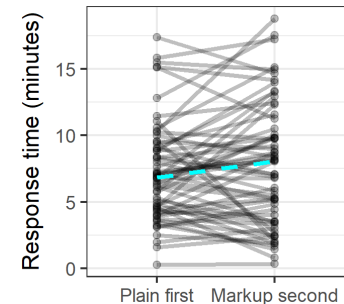
– Trial order

- First trial $>_{ns}$ Second trial (med = 6.64min, 5.93min)
 - 53% of participants (53/100)
 - *No clear learning between first and second trials*



– Condition order

- Plain first: Plain \leq Markup (med = 6.84min, 8.05min)
 - 54% of participants (33/61)
- Plain second: Plain \leq Markup (med = 3.90min, 5.60min)
 - 64% of participants (25/39)
 - *Learning from Markup to Plain?*
(Asymmetric transfer)





RESULTS – 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



RESULTS – 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



RESULTS – 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

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



RESULTS – 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

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



RESULTS – ACCURACY/SPEED BY PREFERENCE

Faster, more accurate in preferred condition?

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



RESULTS – ACCURACY/SPEED BY PREFERENCE

More accurate in preferred condition, faster in Plain?

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

Speed

- Prefer plain $>^*$ Prefer markup (med = 7.06min, 4.48min)

Accuracy

- Prefer plain $>_{ns}$ Prefer markup (med = 5, 4.5)

* Significant by Mann-Whitney U test



RESULTS – PREFERENCE BY FIRST-SEEN CONDITION

Preference for first-seen trial?

Trial order	Preference		
	Plain	Markup	Even
Plain first, markup second	70 2.91	25	5
Markup first, plain second	59 1.51	38	3

Preference for Plain greater
when Plain is seen first



DISCUSSION

Does text markup from an IE pipeline improve human comprehension of text documents?

– Overall

- This markup hurts performance (accuracy, speed)
- Participants associate higher workload with this markup, disprefer

– But

- Minority of participants are more accurate with markup!

– Next steps

- Test “ideal” markup
 - Less, higher quality (accuracy, relevance)



ACKNOWLEDGMENTS AND BIBLIOGRAPHY

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

- J. Nielsen, *Designing web usability: The practice of simplicity*, Indianapolis, IN: New Riders Publishing, 1999.
- 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.
- 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.
- 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.
- 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.
- M. Ruddy, "ELICIT – The Experimental Laboratory for the Investigation of Collaboration, Information Sharing, and Trust," Proceedings of the 12th ICCRTS, Newport, RI, June 2007.
- 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.
- NASA: NASA Task Load Index (TLX), v. 1.0 Manual, 1986.
- E. C. Poulton and P. R. Freeman, "Unwanted asymmetrical transfer effects with balanced experimental designs," in Psychological Bulletin, 1966, pp. 1-8.
- E. T. Davis and J. Palmer, "Visual search and attention: An overview," Spatial Vision, vol 17 (4-5), 2004, pp. 249-255.
- 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.
- 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.

