Learning Script Knowledge with Web Experiments

Michaela Regneri, Alexander Koller, Manfred Pinkal 2010

Presentation by Christian Wellner

Overview

- Motivation
- Experiment 1: Temporal structure
- Experiment 2: Participants
- Conclusion

Motivation

- What is a script?

“a standardized sequence of events that describes some stereotypical human activity such as going to a restaurant or visiting a doctor” (Barr and Feigenbaum, 1981)

Motivation

- A script is usually defined by
  - A set of events
  - Temporal constraints on these events
  - Causal connections between the events
  - Identification of participants of the script

Motivation

- A script is usually defined by
  - A set of events
  - Temporal constraints on these events
  - Causal connections between the events
  - Identification of participants of the script
- „Scenario“ = class of activities vs.

  „Script“ = model of the structure of a scenario

Motivation

- Why could scripts be useful?
  - They provide detailed background knowledge of a given situation

Motivation

- Why could scripts be useful?
  - They provide detailed background knowledge of a given situation
  - Much of this knowledge is implicit and will not be mentioned in texts

Motivation

- Why could scripts be useful?
  - They provide detailed background knowledge of a given situation
  - Much of this knowledge is implicit and will not be mentioned in texts
  - Applications: Anaphora resolution, WSD, information extraction, text generation, ...

Motivation

- Problem: Learn scripts automatically
  - Manual formalization by experts does not scale: Too many variations

Motivation

- Problem: Learn scripts automatically
  - Manual formalization by experts does not scale: Too many variations
  - Standard approach: Learning from corpora (Chambers & Jurafsky 2008)

Motivation

- Problem: Learn scripts automatically
  - Manual formalization by experts does not scale: Too many variations
  - Standard approach: Learning from corpora (Chambers & Jurafsky 2008)
  - Since script knowledge is often highly implicit, it is hard to extract directly from random text sources

Motivation

- Approach of this paper: Ask volunteers for script descriptions

Motivation

- Approach of this paper: Ask volunteers for script descriptions
  - Script descriptions are explicit
  - Domains can be arbitrary
  - Better control over structure / granularity

Experiments

- 2 separate experiments:

Experiments

- 2 separate experiments:
  - Experiment 1: learning events in a script and their temporal structure (and by extension causal structure)

Experiments

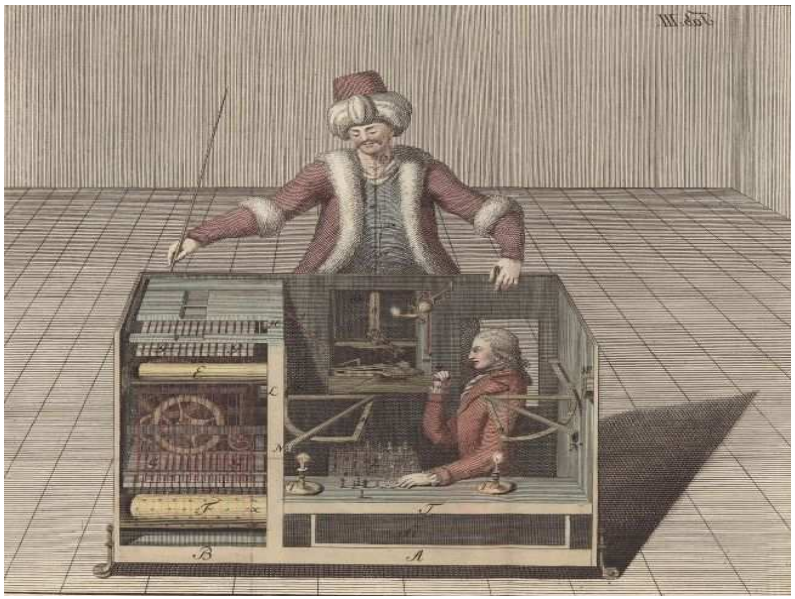
- 2 separate experiments:
  - Experiment 1: learning events in a script and their temporal structure (and by extension causal structure)
  - Experiment 2: learning script participants

Experiment 1

- Data acquisition: via web service „Mechanical Turk“

Experiment 1

- Data acquisition: via web service „Mechanical Turk“



Experiment 1

- Data acquisition
  - Each volunteer is asked to describe some scenario in bullet points

Experiment 1

- Data acquisition
  - Each volunteer is asked to describe some scenario in bullet points
  - One bullet point = natural language realization of an event

Experiment 1

- Data acquisition
  - Each volunteer is asked to describe some scenario in bullet points
  - One bullet point = natural language realization of an event
  - List of events = Event Sequence Description (ESD) → „raw“ script

Experiment 1

- ESDs:

1. look at menu
2. decide what you want
3. order at counter
4. pay at counter
5. receive food at counter
6. take food to table
7. eat food

1. walk to the counter
2. place an order
3. pay the bill
4. wait for the ordered food
5. get the food
6. move to a table
7. eat food
8. exit the place

1. walk into restaurant
2. find the end of the line
3. stand in line
4. look at menu board
5. decide on food and drink
6. tell cashier your order
7. listen to cashier repeat order
8. listen for total price
9. swipe credit card in scanner
10. put up credit card
11. take receipt
12. look at order number
13. take your cup
14. stand off to the side
15. wait for number to be called
16. get your drink

Experiment 1

- First processing step: Calculate Multiple Sequence Alignments (MSA)
  - Algorithm from bioinformatics

Experiment 1

- First processing step: Calculate Multiple Sequence Alignments (MSA)
  - Algorithm from bioinformatics
  - Align ESDs in a matrix

Experiment 1

- First processing step: Calculate Multiple Sequence Alignments (MSA)
  - Algorithm from bioinformatics
  - Align ESDs in a matrix
  - Try to match events of different ESDs

Experiment 1

- First processing step: Calculate Multiple Sequence Alignments (MSA)
  - Algorithm from bioinformatics
  - Align ESDs in a matrix
  - Try to match events of different ESDs
  - If a certain event cannot be matched in a certain ESD, insert a gap (Ø)

Experiment 1

- First processing step: Calculate MSAs

row	s_1	s_2	s_3	s_4
1	⊘	walk into restaurant	⊘	enter restaurant
2	⊘	⊘	walk to the counter	go to counter
3	⊘	find the end of the line	⊘	⊘
4	⊘	stand in line	⊘	⊘
5	look at menu	look at menu board	⊘	⊘
6	decide what you want	decide on food and drink	⊘	make selection
7	order at counter	tell cashier your order	place an order	place order
8	⊘	listen to cashier repeat order	⊘	⊘
9	pay at counter	⊘	pay the bill	pay for food
10	⊘	listen for total price	⊘	⊘
11	⊘	swipe credit card in scanner	⊘	⊘
12	⊘	put up credit card	⊘	⊘
13	⊘	take receipt	⊘	⊘
14	⊘	look at order number	⊘	⊘
15	⊘	take your cup	⊘	⊘
16	⊘	stand off to the side	⊘	⊘
17	⊘	wait for number to be called	wait for the ordered food	⊘
18	receive food at counter	get your drink	get the food	pick up order
19	⊘	⊘	⊘	pick up condiments
20	take food to table	⊘	move to a table	go to table
21	eat food	⊘	eat food	consume food
22	⊘	⊘	⊘	clear tray
22	⊘	⊘	exit the place	⊘

Experiment 1

- First processing step: Calculate MSAs
  - Each MSA A has a cost c(A):

$$c(A) = c_{gap} \cdot \Sigma_{\oslash} + \sum_{i=1}^{n} \sum_{\substack{j=1, \\ a_{ji} \neq \oslash}}^{m} \sum_{\substack{k=j+1, \\ a_{ki} \neq \oslash}}^{m} c_m(a_{ji}, a_{ki})$$

(n = number of rows; m = number of sequences)

Experiment 1

- First processing step: Calculate MSAs
  - Each MSA A has a cost c(A):

$$c(A) = c_{gap} \cdot \Sigma_{\oslash} + \sum_{i=1}^{n} \sum_{\substack{j=1, \\ a_{ji} \neq \oslash}}^{m} \sum_{\substack{k=j+1, \\ a_{ki} \neq \oslash}}^{m} c_m(a_{ji}, a_{ki})$$

(n = number of rows; m = number of sequences)

  - Sum gap costs for each gap plus alignment costs for any two events that are aligned

Experiment 1

- First processing step: Calculate MSAs
  - Cost function c(m): semantic dissimilarity

Experiment 1

- First processing step: Calculate MSAs
  - Cost function c(m): semantic dissimilarity
  - Since the events are in bullet-point-style, traditional bag-of-words does not work

Experiment 1

- First processing step: Calculate MSAs
  - Cost function c(m): semantic dissimilarity
  - Since the events are in bullet-point-style, traditional bag-of-words does not work
  - Heuristic similarity measure:

$$sim = \alpha \cdot pred + \beta \cdot subj + \gamma \cdot obj$$

(α, β, γ : weights; pred, subj and obj are pseudo-parsed)

Experiment 1

- Second processing step: Calculate Temporal Script Graphs (TSG)
  - Directed Graphs with events as nodes

Experiment 1

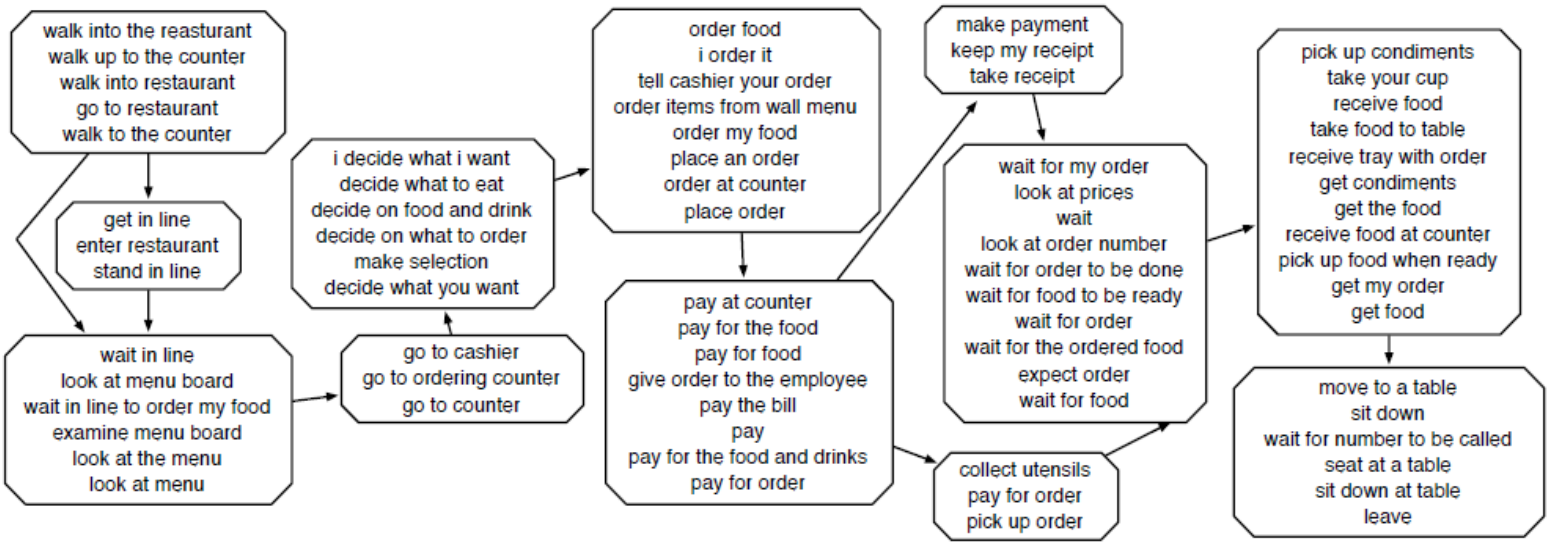
- Second processing step: Calculate Temporal Script Graphs (TSG)
  - Directed Graphs with events as nodes
  - Each row in the MSA → one node

Experiment 1

- Second processing step: Calculate Temporal Script Graphs (TSG)
  - Directed Graphs with events as nodes
  - Each row in the MSA → one node
  - Merge nodes according to structural and semantic constraints

Experiment 1

- Second processing step: Calculate TSGs



Experiment 1

- Evaluation
  - Test for paraphrase recognition and temporal correctness

Experiment 1

- Evaluation
  - Test for paraphrase recognition and temporal correctness
  - Test set: paraphrase/happens-before pairs mixed with random pairs

Experiment 1

- Evaluation
  - Test for paraphrase recognition and temporal correctness
  - Test set: paraphrase/happens-before pairs mixed with random pairs
  - Ask volunteers in Mechanical Turk to rate all pairs → gold standard

Experiment 1

- Evaluation
  - Baseline 1: Clustering of events

Experiment 1

- Evaluation
  - Baseline 1: Clustering of events
  - Baseline 2: Levenshtein measure instead of similiarity function

Experiment 1

- Evaluation
  - Baseline 1: Clustering of events
  - Baseline 2: Levenshtein measure instead of similiarity function
  - + upper bound: human performance

Experiment 1

- Evaluation results(paraphrasing):

	Scenario	Precision			Recall			F-Score			
		sys	base$_{cl}$	base$_{lev}$	*sys*	base$_{cl}$	base$_{lev}$	*sys*	base$_{cl}$	base$_{lev}$	upper
MTurk	pay with credit card	0.52	0.43	0.50	0.84	0.89	0.11	**0.64**	0.58	• 0.17	0.60
	eat in restaurant	0.70	0.42	0.75	0.88	1.00	0.25	**0.78**	• 0.59	• 0.38	• 0.92
	iron clothes I	0.52	0.32	1.00	0.94	1.00	0.12	**0.67**	• 0.48	• 0.21	• 0.82
	cook scrambled eggs	0.58	0.34	0.50	0.86	0.95	0.10	**0.69**	• 0.50	• 0.16	• 0.91
	take a bus	0.65	0.42	0.40	0.87	1.00	0.09	**0.74**	• 0.59	• 0.14	• 0.88
OMICS	answer the phone	0.93	0.45	0.70	0.85	1.00	0.21	**0.89**	• 0.71	• 0.33	0.79
	buy from vending machine	0.59	0.43	0.59	0.83	1.00	0.54	**0.69**	0.60	0.57	0.80
	iron clothes II	0.57	0.30	0.33	0.94	1.00	0.22	**0.71**	• 0.46	• 0.27	0.77
	make coffee	0.50	0.27	0.56	0.94	1.00	0.31	**0.65**	• 0.42	○ 0.40	• 0.82
	make omelette	0.75	0.54	0.67	0.92	0.96	0.23	**0.83**	• 0.69	• 0.34	0.85
	Average	0.63	0.40	0.60	0.89	0.98	0.22	**0.73**	0.56	0.30	0.82

Experiment 1

- Evaluation results(paraphrasing):
  - System outperforms baselines
  - Sometimes close to upper bound

Experiment 1

- Evaluation results(paraphrasing):
  - System outperforms baselines
  - Sometimes close to upper bound
  - Clustering has problems distinguishing similiarities (high recall, low precision)

Experiment 1

- Evaluation results(paraphrasing):
  - System outperforms baselines
  - Sometimes close to upper bound
  - Clustering has problems distinguishing similiarities (high recall, low precision)
  - Levenshtein is too restrictive (low recall, high precision)

Experiment 1

- Evaluation results(happens-before):

	Scenario	Precision			Recall			F-Score			
		sys	base$_{cl}$	base$_{lev}$	*sys*	base$_{cl}$	base$_{lev}$	*sys*	base$_{cl}$	base$_{lev}$	upper
MTurk	pay with credit card	0.86	0.49	0.65	0.84	0.74	0.45	**0.85**	• 0.59	• 0.53	0.92
	eat in restaurant	0.78	0.48	0.68	0.84	0.98	0.75	**0.81**	• 0.64	0.71	• 0.95
	iron clothes I	0.78	0.54	0.75	0.72	0.95	0.53	**0.75**	0.69	• 0.62	• 0.92
	cook scrambled eggs	0.67	0.54	0.55	0.64	0.98	0.69	0.66	***0.70***	0.61	• 0.88
	take a bus	0.80	0.49	0.68	0.80	1.00	0.37	**0.80**	• 0.66	• 0.48	• 0.96
OMICS	answer the phone	0.83	0.48	0.79	0.86	1.00	0.96	0.84	• 0.64	***0.87***	0.90
	buy from vending machine	0.84	0.51	0.69	0.85	0.90	0.75	**0.84**	• 0.66	○ 0.71	0.83
	iron clothes II	0.78	0.48	0.75	0.80	0.96	0.66	**0.79**	• 0.64	0.70	0.84
	make coffee	0.70	0.55	0.50	0.78	1.00	0.55	**0.74**	0.71	○ 0.53	○ 0.83
	make omelette	0.70	0.55	0.79	0.83	0.93	0.82	0.76	○ 0.69	***0.81***	• 0.92
	Average	0.77	0.51	0.68	0.80	0.95	0.65	**0.78**	0.66	0.66	0.90

Experiment 1

- Evaluation results(happens-before):
  - Basically the same as for paraphrases

Experiment 1

- Evaluation results(happens-before):
  - Basically the same as for paraphrases
  - Higher average because temporal ordering is easier once the paraphrases are correct

Experiment 1

- Evaluation results(happens-before):
  - Basically the same as for paraphrases
  - Higher average because temporal ordering is easier once the paraphrases are correct
  - More problems with scripts that are flexible in ordering (some events may happen in arbitrary order)

Experiment 2

- Data: based on TSGs of Experiment 1

Experiment 2

- Data: based on TSGs of Experiment 1
- Goal: Identify participants in a script

Experiment 2

- Data: based on TSGs of Experiment 1
- Goal: Identify participants in a script
- Method: Identify possible participants, then run an Integer Linear Program (ILP) to sort them into equivalence classes

Experiment 2

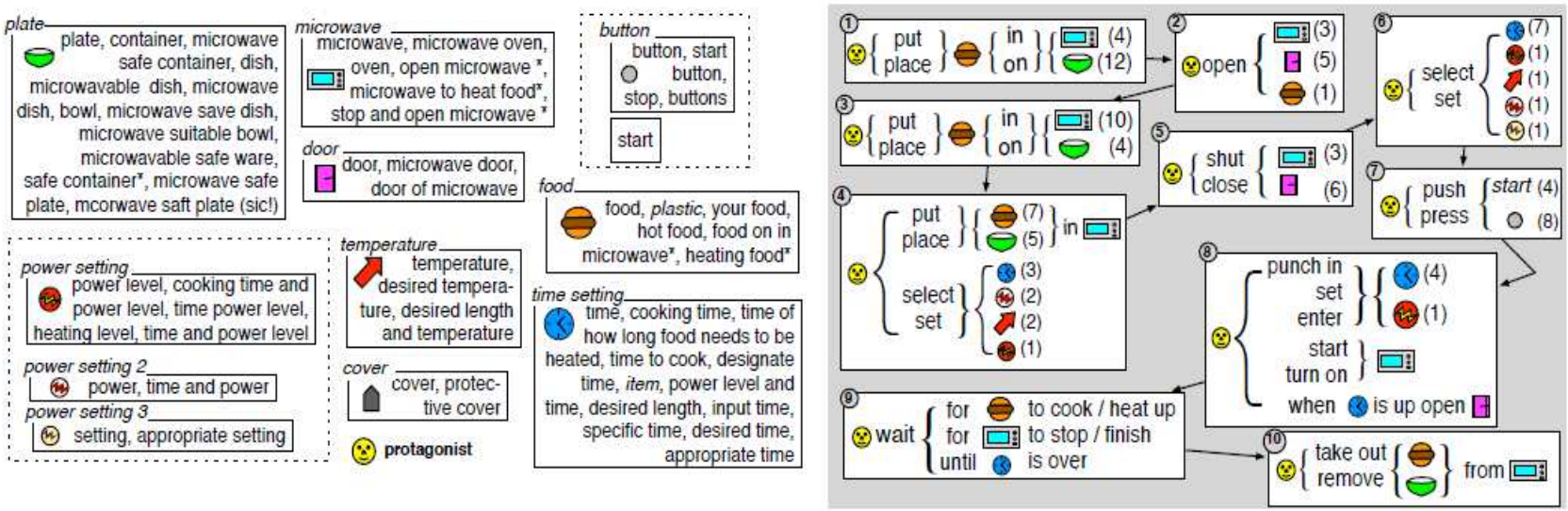


Figure 4: The participants we extracted for the MICROWAVE scenario, and a participant-annotated excerpt from the original graph. Descriptions in *italics* indicate sorting mistakes, asterisks (*) mark parsing mistakes. Dotted boxes frame PDSs that actually belong together but were not combined by the algorithm.

Experiment 2

- Evaluation:
  - Gold standard by annotators

Experiment 2

- Evaluation:
  - Gold standard by annotators
  - Baseline 1: grouping by string equivalence

Experiment 2

- Evaluation:
  - Gold standard by annotators
  - Baseline 1: grouping by string equivalence
  - Baseline 2: grouping by semantic similarity

Experiment 2

- Evaluation:
  - Gold standard by annotators
  - Baseline 1: grouping by string equivalence
  - Baseline 2: grouping by semantic similarity
  - Baseline 3: grouping by structural similiarity

Experiment 2

- Evaluation results:

Scenario	Precision				Recall				F-Score			
	full	sem	align	base	full	sem	align	base	full	sem	align	base
LAUNDRY*	0.85	0.76	0.53	0.93	0.75	0.83	0.89	0.57	**0.80**	0.79	0.67	0.70
VENDING M.*	0.80	0.74	0.57	0.84	0.78	0.83	0.97	0.62	**0.79**	0.78	0.72	0.72
FAST FOOD	0.82	0.65	0.55	0.87	0.82	0.85	0.84	0.70	**0.82**	0.74	0.66	0.78
RETURN FOOD	0.80	0.78	0.53	0.88	0.44	0.52	0.63	0.34	0.57	***0.62***	0.57	0.49
COFFEE	0.85	0.77	0.53	0.92	0.80	0.81	0.98	0.68	**0.82**	0.79	0.68	0.78
FEED DOG	0.81	0.67	0.53	0.90	0.88	0.92	0.94	0.57	**0.84**	0.78	0.68	0.70
MICROWAVE	0.89	0.78	0.55	0.93	0.84	0.84	0.89	0.70	**0.86**	0.81	0.68	0.80
CREDIT CARD	0.90	0.82	0.60	0.94	0.54	0.54	0.64	0.40	**0.67**	0.65	0.62	0.56
MAIL LETTER	0.92	0.78	0.54	0.96	0.88	0.88	0.93	0.74	**0.90**	0.83	0.68	0.84
SHOWER	0.87	0.79	0.57	0.94	0.83	0.83	0.86	0.66	**0.85**	0.81	0.69	0.77
AVERAGE*	0.85	0.75	0.55	0.91	0.75	0.79	0.86	0.60	• **0.79**	• 0.76	0.66	0.71
AVERAGE	0.86	0.76	0.55	0.92	0.75	0.77	0.84	0.60	• **0.79**	0.75	0.66	0.71

Experiment 2

- Evaluation results:
  - sem scores high precision, low recall

Experiment 2

- Evaluation results:
  - sem scores high precision, low recall
  - align scores high recall, low precision

Experiment 2

- Evaluation results:
  - sem scores high precision, low recall
  - align scores high recall, low precision
  - Full system scores a better trade-off that leads to a better F-Score

Conclusion

- The presented system yields better results than the baselines

Conclusion

- The presented system yields better results than the baselines
- Sometimes it reaches human performance

Conclusion

- The presented system yields better results than the baselines
- Sometimes it reaches human performance
- Some fine-tuning may be possible

Conclusion

- The presented system yields better results than the baselines
- Sometimes it reaches human performance
- Some fine-tuning may be possible
- System may perform worse for culture-specific or expert knowledge (depending on volunteers)

Conclusion

- Remaining problem: decrease supervision
  - Scenarios were still chosen manually

Conclusion

- Remaining problem: decrease supervision
  - Scenarios were still chosen manually
  - Idea: use online games to generate new candidate scenarios

Conclusion

- Remaining problem: decrease supervision
  - Scenarios were still chosen manually
  - Idea: use online games to generate new candidate scenarios
  - Filter raw data automatically

Conclusion

- Remaining problem: decrease supervision
  - Scenarios were still chosen manually
  - Idea: use online games to generate new candidate scenarios
  - Filter raw data automatically
  - Restrict user input to reduce need for orthography correction

Thank you for your attention!