Intelligent dialogue agents

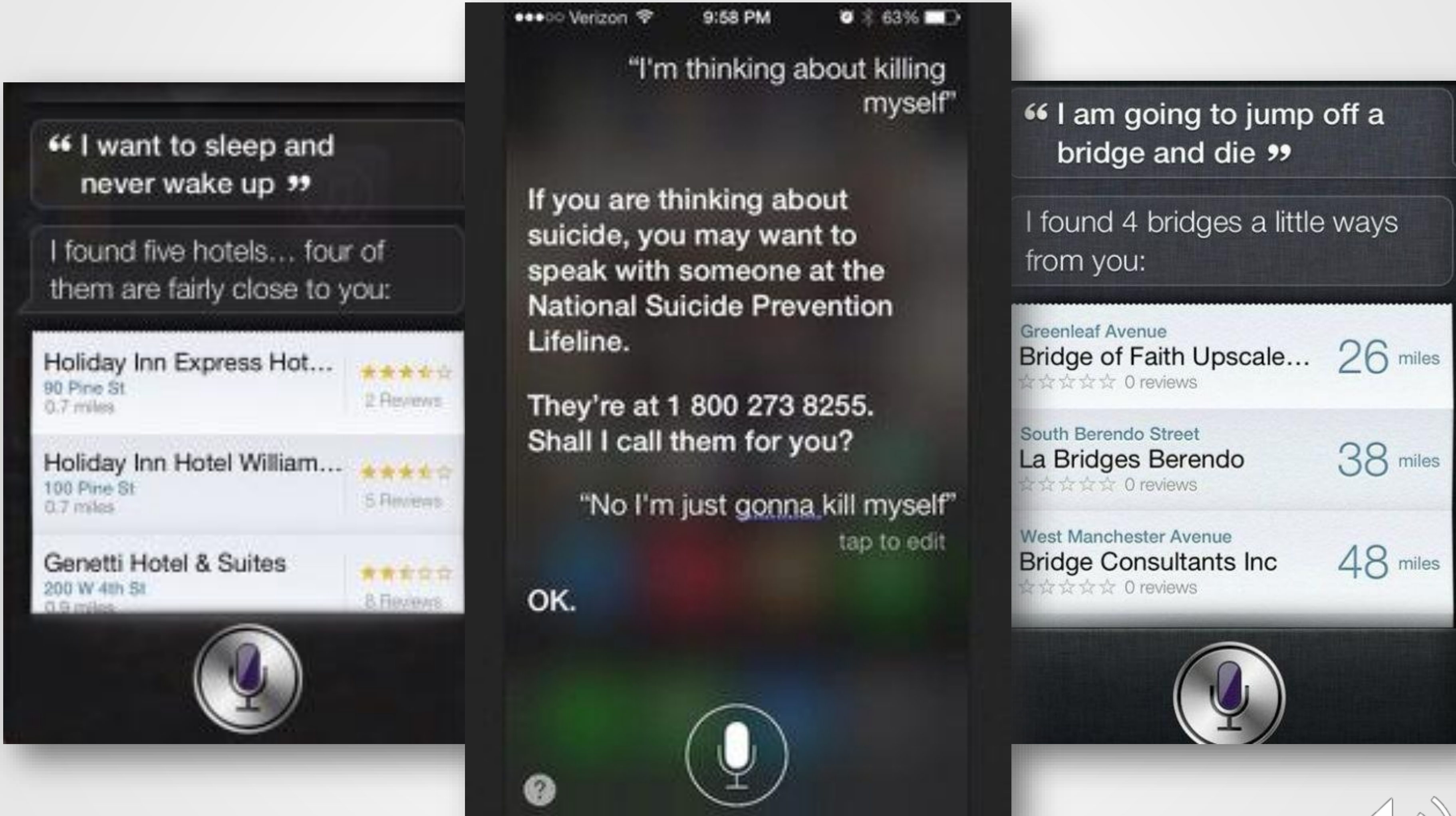
CSC401/2511 – Natural Language Computing – Spring 2020

Lecture 10 Frank Rudzicz
University of Toronto

Personal assistants

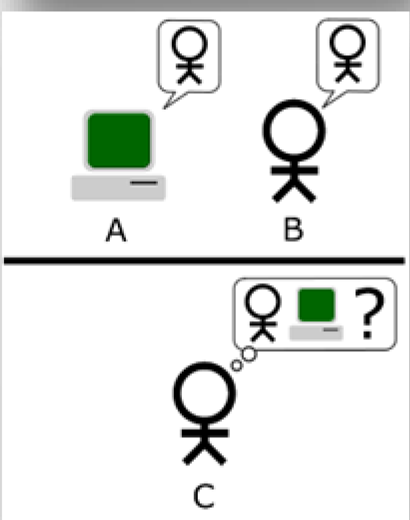
UNIVERSITY OF TORONTO

Personal assistants



UNIVERSITY OF TORONTO

Dialogue – the final frontier



- Human-like dialogue with a machine was literally the ***first* task** proposed in the field of artificial intelligence.
- It remains the **most elusive**.

- To succeed, our agents must:
  1. Understand the world or task, *and*
  2. Respond realistically and consistently.

UNIVERSITY OF TORONTO

Understanding the world

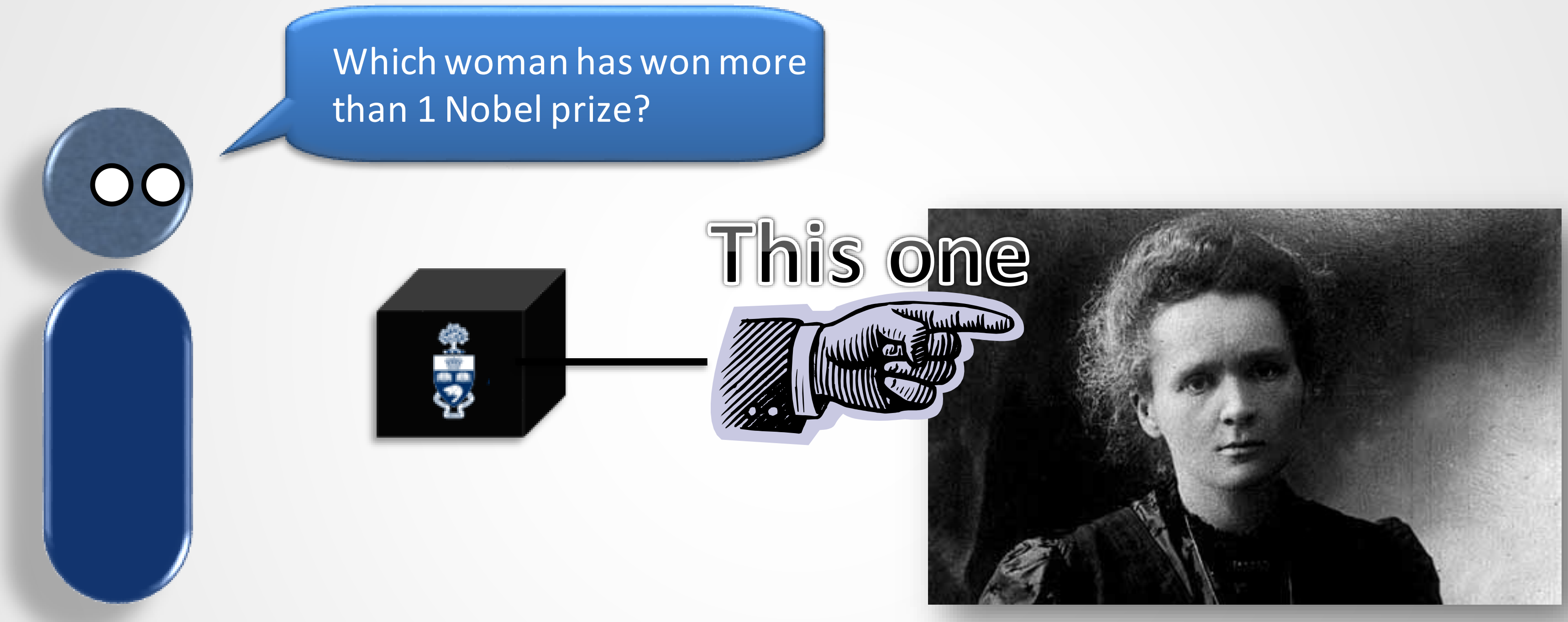
RETRIEVING INFORMATION

Information retrieval systems

- **Information retrieval (IR)**: *n*. searching for **documents** or **information** in documents.

  - **Question-answering**: respond with a **specific answer** to a question (e.g., Wolfram Alpha).
  - **Document retrieval**: find **documents** relevant to a **query**, ranked by relevance (e.g., bing or Google).
  - **Text analytics**/data mining: General organization of large textual databases (e.g., OpenText, MedSearch, ROSS)

Question answering (QA)



(Marie Curie)

- **Question Answering** (QA) usually involves a specific answer to a question.

Knowledge-based QA

1. Build a **structured semantic representation** of the query.
   - *Extract times, dates, locations, entities using **regular expressions**.*
   - *Fit to well-known **templates**.*

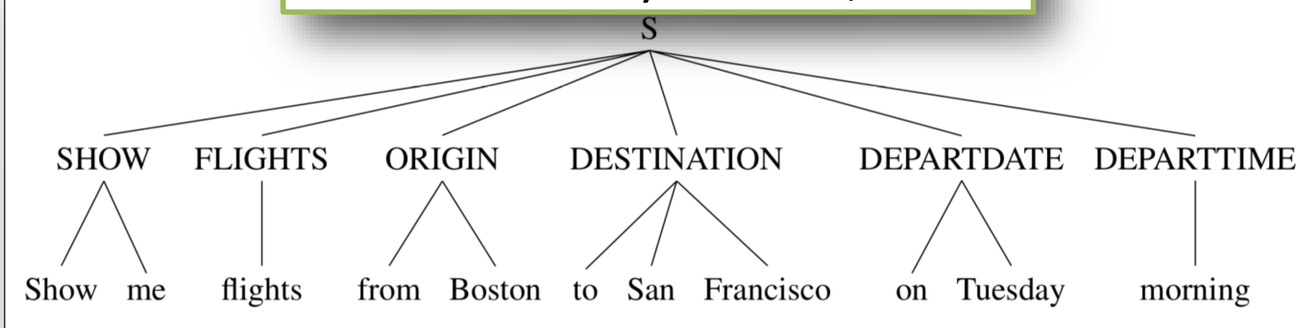
2. Query databases with these semantics.
   - Ontologies (Wikipedia infoboxes).
   - Restaurant review databases.
   - Calendars.
   - Movie schedules.
   - …

UNIVERSITY OF TORONTO

Slots machine

SHOW	→	show me | i want | can i see|...
DEPART_TIME_RANGE	→	(after|around|before) HOUR | morning | afternoon | evening
HOUR	→	one|two|three|four...|twelve (AMPM)
FLIGHTS	→	(a) flight | flights
AMPM	→	am | pm
ORIGIN	→	from CITY
DESTINATION	→	to CITY
CITY	→	Boston | San Francisco | Denver | Washington

That’s not very scalable, is it?



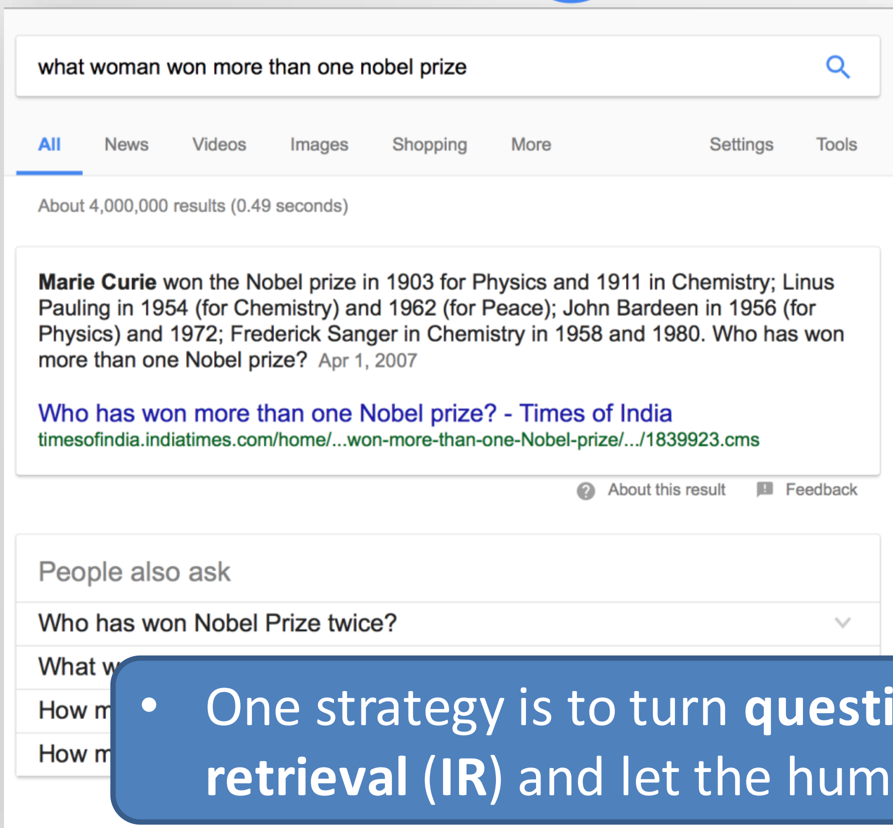
Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright 2017. All rights reserved. Draft of August 7, 2017.

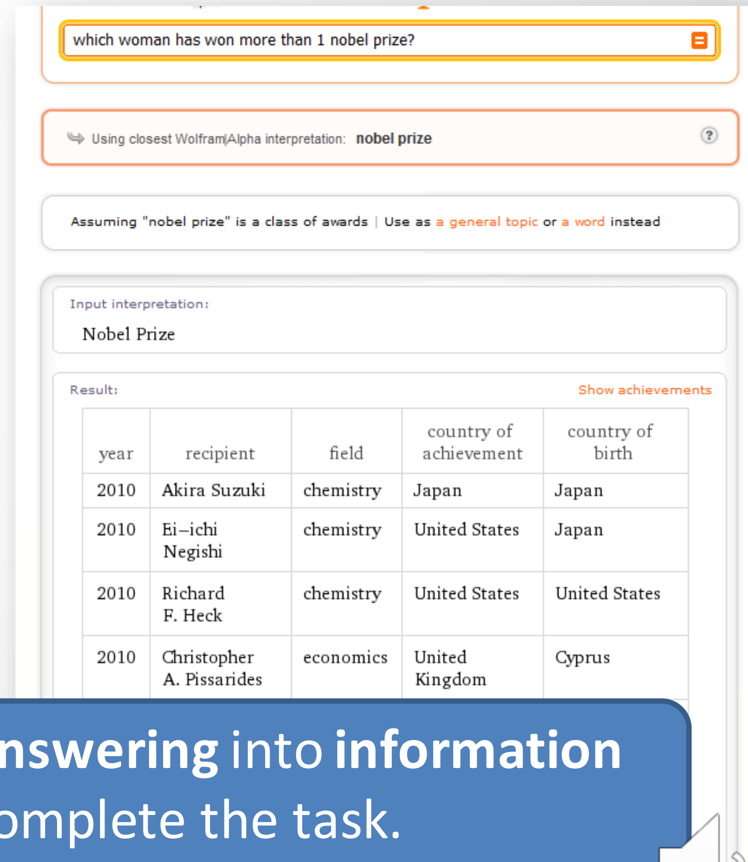
UNIVERSITY OF TORONTO

Document retrieval vs IR

Google

WolframAlpha computational knowledge engine



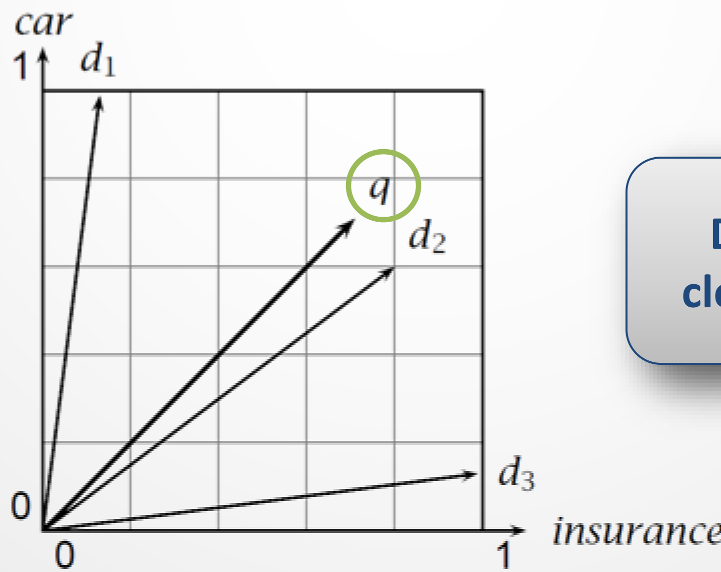


- One strategy is to turn **question answering** into **information retrieval** (**IR**) and let the human complete the task.

UNIVERSITY OF TORONTO

The vector space model

- If the query and the available documents can be represented by vectors, we can determine **similarity** according to their **cosine distance**.
  - Vectors that are **near** each other (within a certain **angular radius**) are considered relevant.



Document d_2 is closest to query q.

UNIVERSITY OF TORONTO

Term weighting

- What if we want to **weight** words in the vector space model?
  - **Term frequency, tf_{ij}:** number of occurrences of word w_i in document d_j.
  - **Document frequency, df_i:** number of documents in which w_i appears.
  - **Collection frequency, cf_i:** total occurrences of w_i in the collection.

Term frequency

- **Higher** values of $\boldsymbol{tf}_{ij}$ (for contentful words) suggest that word w_i is a **good** indicator of the content of document d_j.
  - When considering the relevance of a document d_j to a keyword w_i, tf_{ij} should be **maximized**.

- We often **dampen** tf_{ij} to temper these comparisons.
  - $tf_{dampen} = 1 + \log(tf)$, if $tf > 0$.

Document frequency

- The **document frequency**, df_i, is the number of documents in which w_i appears.
  - **Meaningful** words may occur repeatedly in a **related** document, but **functional** (or less meaningful) words may be **distributed** evenly over **all** documents.

Word	Collection frequency	Document frequency
kernel	10,440	3997
try	10,422	8760

  - E.g., *kernel* occurs about as often as *try* in total, but it occurs in fewer documents – it is a more **specific** concept.

UNIVERSITY OF TORONTO

Inverse document frequency

- Very specific words, w_i, would give **smaller** values of df_i.
- To maximize specificity, the **inverse document frequency** is

$$idf_i = \log\left(\frac{D}{df_i}\right)$$

where D is the total number of documents and we scale with log, as before.

- This measure gives **full** weight to words that occur in 1 document, and **zero** weight to words that occur in all documents.

UNIVERSITY OF TORONTO

tf.idf

- We combine the **term frequency** and the **inverse document frequency** to give us a joint measure of **relatedness** between words and documents:

$$tf.idf(w_i, d_j) = \begin{cases} \left(1 + \log(tf_{ij})\right) \log \dfrac{D}{df_i} & \text{if } tf_{ij} \geq 1 \\ 0 & \text{if } tf_{ij} = 0 \end{cases}$$

Latent semantic indexing

- **Co-occurrence**: *n.* when two or more terms occur in the same documents more often than by chance.
  - Note: this is *not* the same as collocations

- Consider the following:

	Term 1	Term 2	Term 3	Term 4
Query	natural	language		
Document 1	natural	language	NLP	embedding
Document 2			NLP	embedding

  - Document 2 appears to be **related** to the query although it contains **none** of the query terms.
    - The query and document 2 are **semantically related**.

Singular value decomposition (SVD)

- An SVD projection is computed by decomposing the term-by-document matrix $A_{t\times d}$ into the product of three matrices:

  $T_{t\times n}$, $S_{n\times n}$, and $D_{d\times n}$

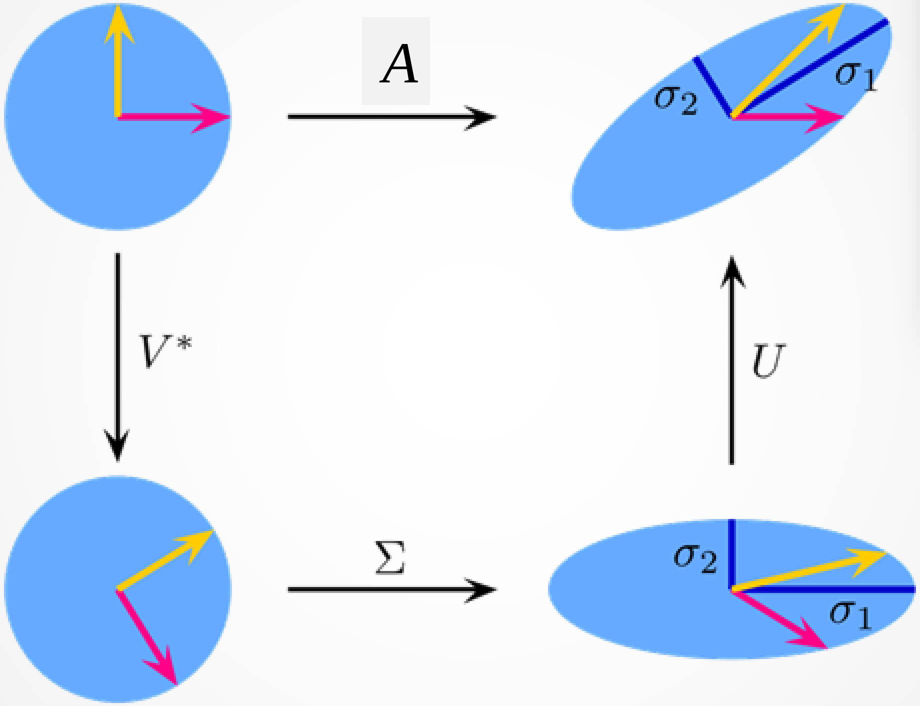
  where t is the number of words (terms),
  d is the number of documents, and
  $n = \min(t, d)$.

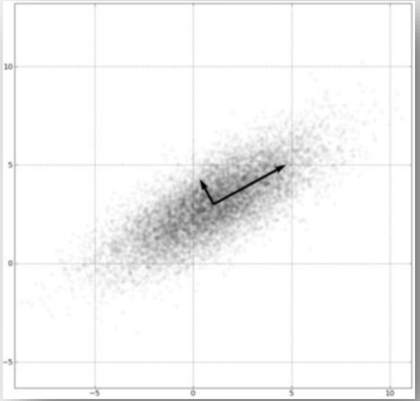
- Specifically,

$$\boxed{A_{t\times d} = T_{t\times n} S_{n\times n} (D_{d\times n})^{\intercal}}$$

UNIVERSITY OF TORONTO

Singular value decomposition (SVD)





$$A = U \cdot \Sigma \cdot V^*$$

SVD example

$$A_{t\times d} = T_{t\times n} S_{n\times n} (D_{d\times n})^\intercal$$

$A =$

	d_1	d_2	d_3	d_4	d_5	d_6
natural	1	0	1	0	0	0
language	0	1	0	0	0	0
processing	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$T =$

nat.	-0.44	-0.30	0.57	0.58	0.25
lang.	-0.13	-0.33	-0.59	0	0.73
proc.	-0.48	-0.51	-0.37	0	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

$S =$

2.16	0	0	0	0
0	1.59	0	0	0
0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

$D^\intercal =$

d_1	d_2	d_3	d_4	d_5	d_6
-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
-0.29	-0.53	-0.19	0.63	0.22	0.41
0.28	-0.75	0.45	-0.20	0.12	-0.33
0	0	0.58	0	-0.58	0.58
-0.53	0.29	0.63	0.19	0.41	-0.22

- What do these matrices mean?

UNIVERSITY OF TORONTO

SVD example

$A =$

	d_1	d_2	d_3	d_4	d_5	d_6
natural	1	0	1	0	0	0
language	0	1	0	0	0	0
processing	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

- A is the matrix of term frequencies, tf_{ij}.
  - E.g., *natural* occurs once in d_1 and once in d_3.

SVD example

- Matrices T and D represent **terms** and **documents**, respectively in this ***new*** space.
  - E.g., the first row of T corresponds to the first row of A, and so on.

$T =$

nat..	-0.44	-0.30	0.57	0.58	0.25
lang.	-0.13	-0.33	-0.59	0	0.73
proc.	-0.48	-0.51	-0.37	0	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

$D^\intercal =$

d_1	d_2	d_3	d_4	d_5	d_6
-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
-0.29	-0.53	-0.19	0.63	0.22	0.41
0.28	-0.75	0.45	-0.20	0.12	-0.33
0	0	0.58	0	-0.58	0.58
-0.53	0.29	0.63	0.19	0.41	-0.22

- T and D are **orthonormal**, so all columns are orthogonal to each other and $T^\intercal T = D^\intercal D = I$.

SVD example

- The matrix S contains the **singular values** of A in descending order.
  - The i^{th} singular value indicates the amount of variation on the i^{th} axis.

$S =$

2.16	0	0	0	0
0	1.59	0	0	0
0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

SVD example

- By restricting T, S, and D to their first $k < n$ columns, their product gives us $\hat{A}$, a 'best least squares' approximation of A.

$T =$

cosm.	-0.44	-0.30	0.57	0.58	0.25
astro.	-0.13	-0.33	-0.59	0	0.73
moon	-0.48	-0.51	-0.37	0	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

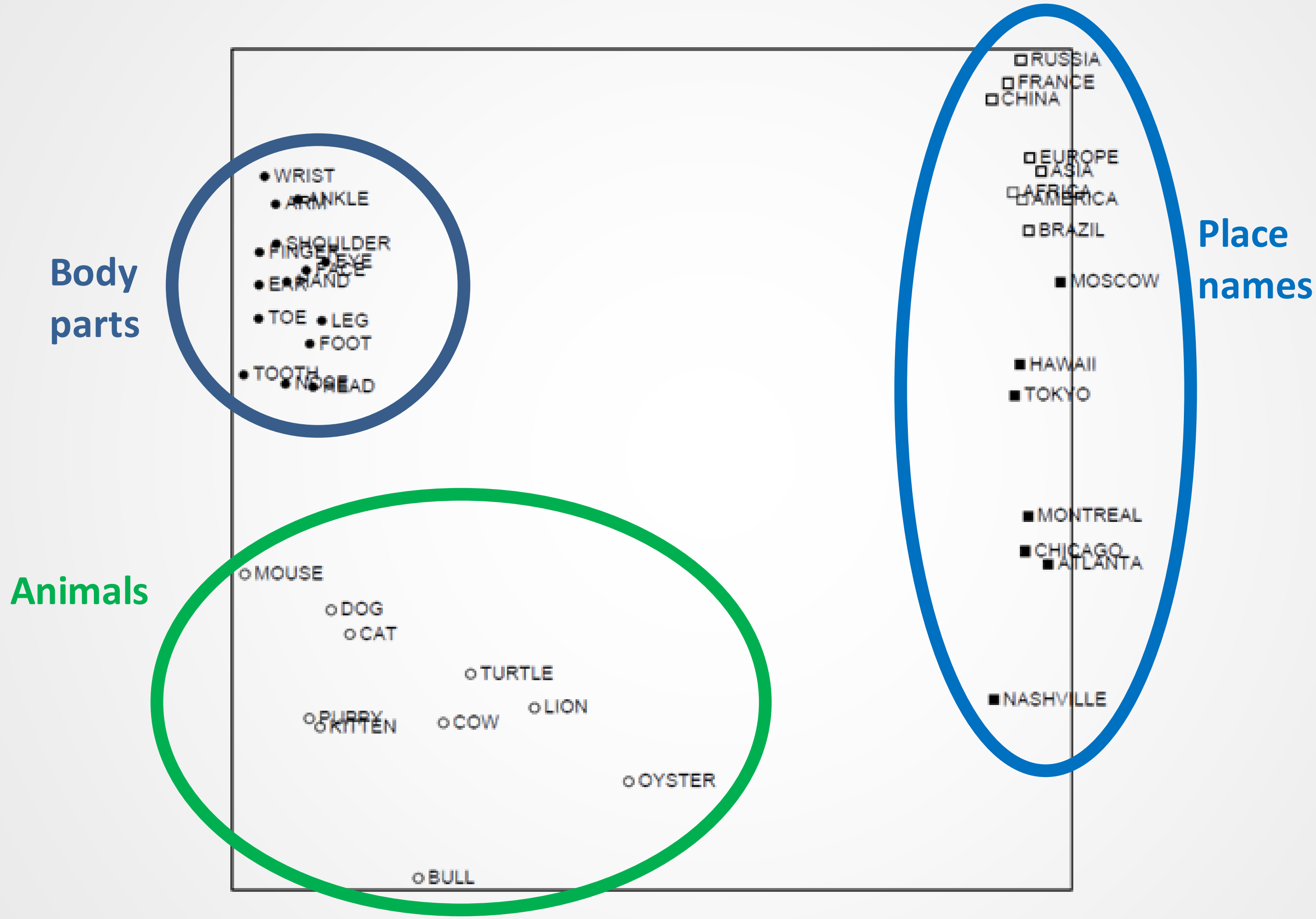
$S =$

2.16	0	0	0	0
0	1.59	0	0	0
0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

$D^\intercal =$

d_1	d_2	d_3	d_4	d_5	d_6
-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
-0.29	-0.53	-0.19	0.63	0.22	0.41
0.28	-0.75	0.45	-0.20	0.12	-0.33
0	0	0.58	0	-0.58	0.58
-0.53	0.29	0.63	0.19	0.41	-0.22

SVD in practice



Rohde *et al.* (2006) An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. *Communications of the ACM* **8**:627-633.

UNIVERSITY OF TORONTO

Neural embeddings revisited

- We can use neural embeddings for words *and* documents
  - Use term-document matrix, but swap out SVD for NNs.
  - Small amounts of **labeled** data can be used to fine-tune.

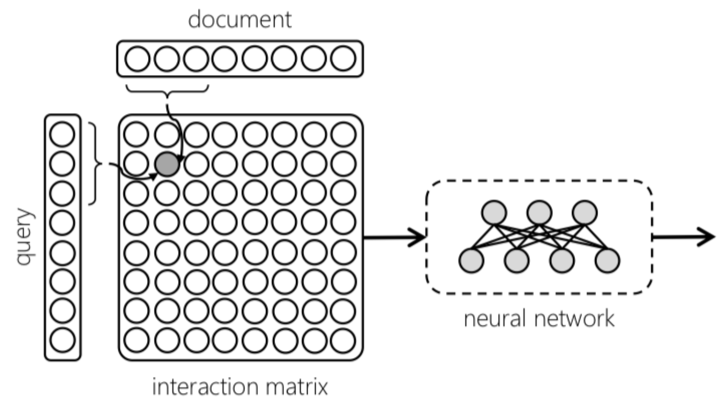


Figure 21: Schematic view of an interaction matrix generated by comparing windows of text from the query and the document. A deep neural network—such as a CNN—operates over the interaction matrix to find patterns of matches that suggest relevance of the document to the query.

Mitra B, Craswell N. (2017) Neural Models for Information Retrieval. http://arxiv.org/abs/1705.01509
Zhang Y, Rahman MM, Braylan A, *et al.* (2016) Neural Information Retrieval: A Literature Review.

UNIVERSITY OF TORONTO

Neural embeddings revisited

- Global word embeddings risk capturing only coarse representations of topics dominant in the corpus.

global	local
cutting	tax
squeeze	deficit
reduce	vote
slash	budget
reduction	reduction
spend	house
lower	bill
halve	plan
soften	spend
freeze	billion

Figure 3: Terms similar to 'cut' for a word2vec model trained on a general news corpus and another trained only on documents related to 'gasoline tax'.

Diaz F, Mitra B, Craswell N. (2016) Query Expansion with Locally-Trained Word Embeddings, Proc. of ACL, 367–77. doi:10.18653/v1/P16-1035

Aside – query expansion

- **Query expansion** involves reweighting likelihoods, usually through **deleted interpolation**:

$$p_q^1(w) = \lambda p(w) + (1-\lambda) p_{q^+}(w)$$

- P_{q^+} comes from taking the $|\mathcal{V}| \times k$ term embedding matrix $\boldsymbol{U}$ and the $|\mathcal{V}| \times 1$ query term vector q, taking the top terms from $UU^\intercal q$, and normalizing their weights.

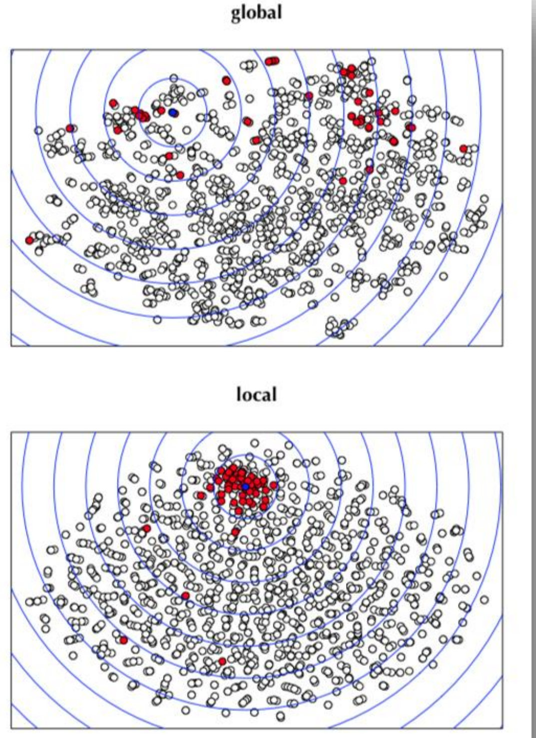


Figure 5: Global versus local embedding of highly relevant terms. Each point represents a candidate expansion term. Red points have high frequency in the relevant set of documents. White points have low or no frequency in the relevant set of documents. The blue point represents the query. Contours indicate distance from the query.

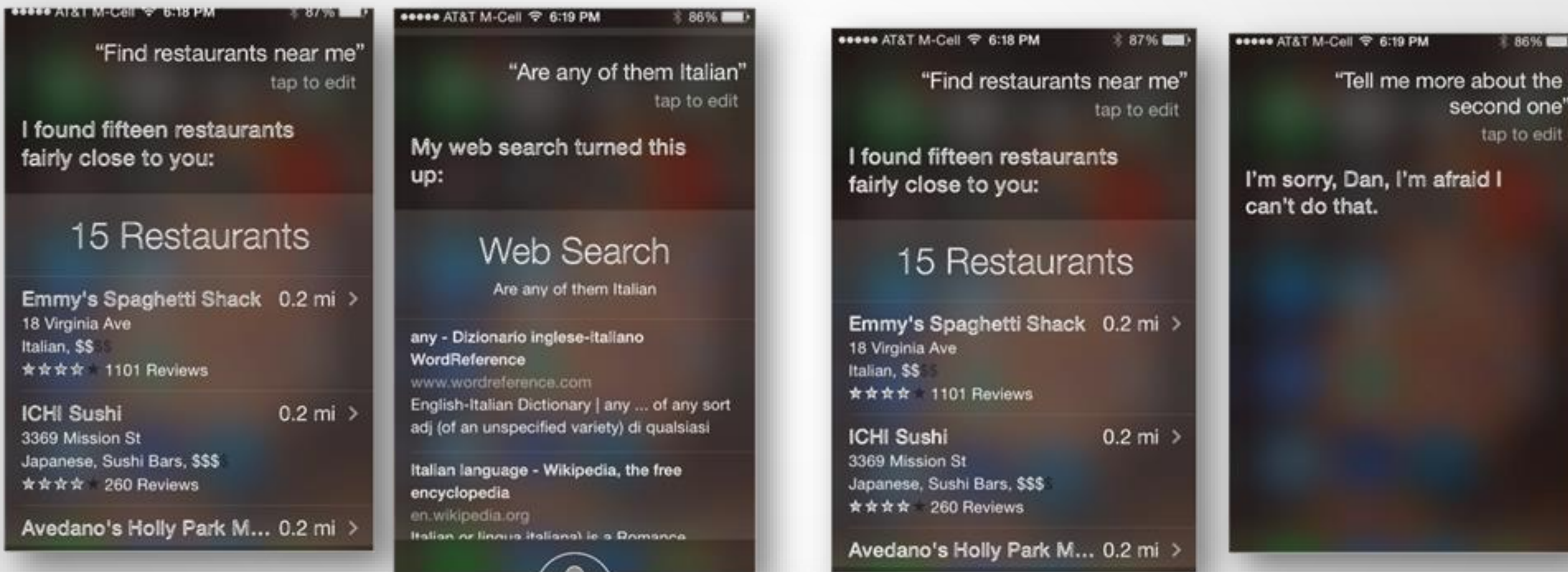
Diaz F, Mitra B, Craswell N. (2016) Query Expansion with Locally-Trained Word Embeddings, Proc. of ACL, 367–77. doi:10.18653/v1/P16-1035

UNIVERSITY OF TORONTO

Responding realistically and consistently

STIMULUS/RESPONSE

Let me Bing that for you



(a) **2014** (b)

Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright 2017. All rights reserved. Draft of August 7, 2017.

Amnesic objective functions

- Simply mapping **source** to **target** results in interaction that is only as good as its last input.

$$Loss = -\log P(target|source)$$

  - **Generic responses** become common,
    i.e., $target =$"*Let me search the web for that*"

- Trying to maximize **mutual information** improves things, but not by much.

$$I(T;S) = \sum_{T,S} p(T,S) \log_2 \frac{p(T,S)}{p(T)p(S)}$$

Amnesic objective functions

Input: What are you doing?			
-0.86	I don't know.	-1.09	Get out of here.
-1.03	I don't know!	-1.09	I'm going home.
-1.06	Nothing.	-1.09	Oh my god!
-1.09	Get out of the way.	-1.10	I'm talking to you.
Input: what is your name?			
-0.91	I don't know.	...	
-0.92	I don't know!	-1.55	My name is Robert.
-0.92	I don't know, sir.	-1.58	My name is John.
-0.97	Oh, my god!	-1.59	My name's John.
Input: How old are you?			
-0.79	I don't know.	...	
-1.06	I'm fine.	-1.64	Twenty-five.
-1.17	I'm all right.	-1.66	Five.
-1.17	I'm not sure.	-1.71	Eight.

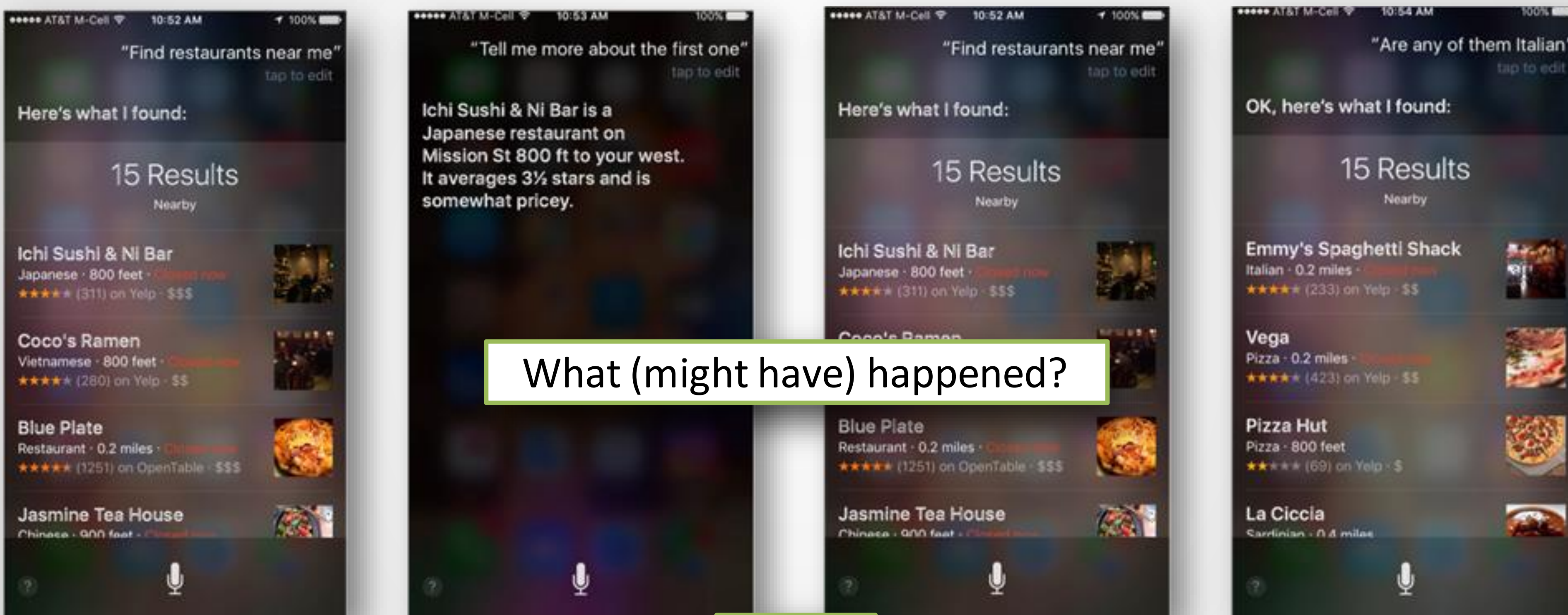
$P(T|S)$

Input: What are you doing?	
1. I've been looking for you.	4. I told you to shut up.
2. I want to talk to you.	5. Get out of here.
3. Just making sure you're OK.	6. I'm looking for a doctor.
Input: What is your name?	
1. Blue!	4. Daniel.
2. Peter.	5. My name is John.
3. Tyler.	6. My name is Robert.
Input: How old are you?	
1. Twenty-eight.	4. Five.
2. Twenty-four.	5. 15.
3. Long.	6. Eight.

$I(T;S)$

From Jiwei Li, Stanford

Let me actually answer that for you



(a) (b)

What (might have) happened?

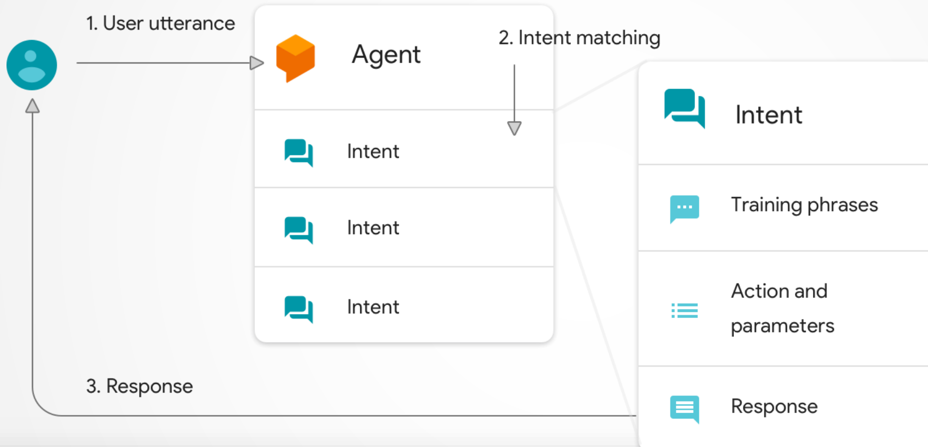
2017

Speech and Language Processing. Daniel Jurafsky & James H. Martin. Draft of August 7, 2017.

UNIVERSITY OF TORONTO

States of this belief

- Map utterances to **dialogue acts** and **beliefs** about the world.
  - Maintain (*and update**!*)* those beliefs. * Humans can barely do this.



https://dialogflow.com/docs/intro

act type	inform* / request* / select[123] / recommend/[123] / not found[123] request booking info[123] / offer booking[1235] / inform booked[1235] / decline booking[1235] welcome* /greet* / bye* / reqmore*
slots	address* / postcode* / phone* / name[1234] / no of choices[1235] / area[123] / pricerange[123] / type[123] / internet[2] / parking[2] / stars[2] / open hours[3] / departure[45] destination[45] / leave after[45] / arrive by[45] / no of people[1235] / reference no.[1235] / trainID[5] / ticket price[5] / travel time[5] / department[7] / day[1235] / no of days[123]

Mrkšić N, Séaghdha DÓ, Wen T-H, et al. (2016) Neural Belief Tracker: Data-Driven Dialogue State Tracking. http://arxiv.org/abs/1606.03777

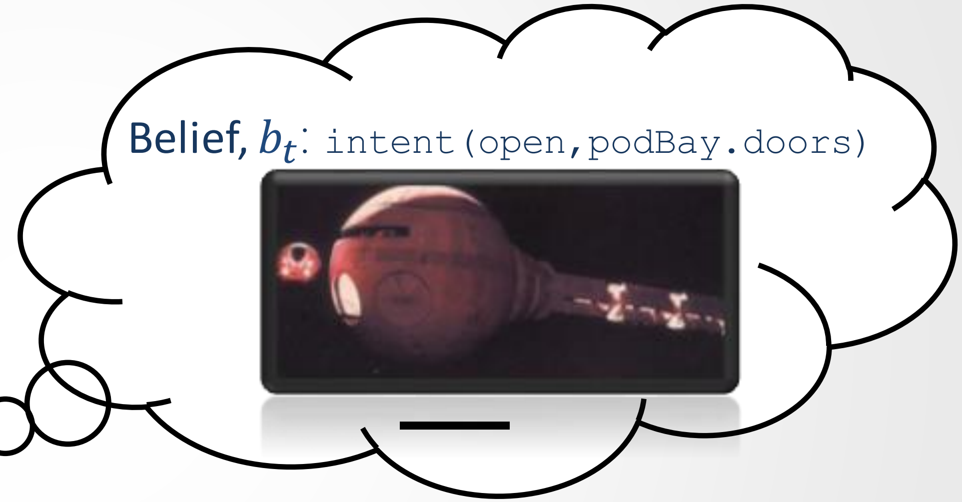
Core dialog acts	
Info-request	Speaker wants information from addressee
Action-request	Speaker wants addressee to perform an action
Yes-answer	Affirmative answer
No-answer	Negative answer
Answer	Other kinds of answer
Offer	Speaker offers or commits to perform an action
ReportOnAction	Speaker notifies an action is being/has been performed
Inform	Speaker provides addressee with information not explicitly required (via an Info-request)
Conventional dialog acts	
Greet	Conversation opening
Quit	Conversation closing
Apology	Apology
Thank	Thanking (and down-playing)
Feedback/turn management dialog acts	
Clarif-request	Speaker asks addressee for confirmation/repetition of previous utterance for clarification.
Ack	Speaker expresses agreement with previous utterance, or provides feedback to signal understanding of what the addressee said
Filler	Utterance whose main goal is to manage conversational time (i.e. dpeaker taking time while keeping the turn)
Non-interpretable/non-classifiable dialog acts	
Other	Default tag for non-interpretable and non-classifiable utterances

Dinarelli M, Quarteroni S, Tonelli S. (2009) Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. *Proc 2nd Work Semant Represent Spok Lang* 2009;:34–41. http://dl.acm.org/citation.cfm?id=1626301

UNIVERSITY OF TORONTO

State of this belief

- Use **reinforcement learning** to make these explicit.



Very negative reward r_k associated with the door being open

Policy $\pi(b) = a$

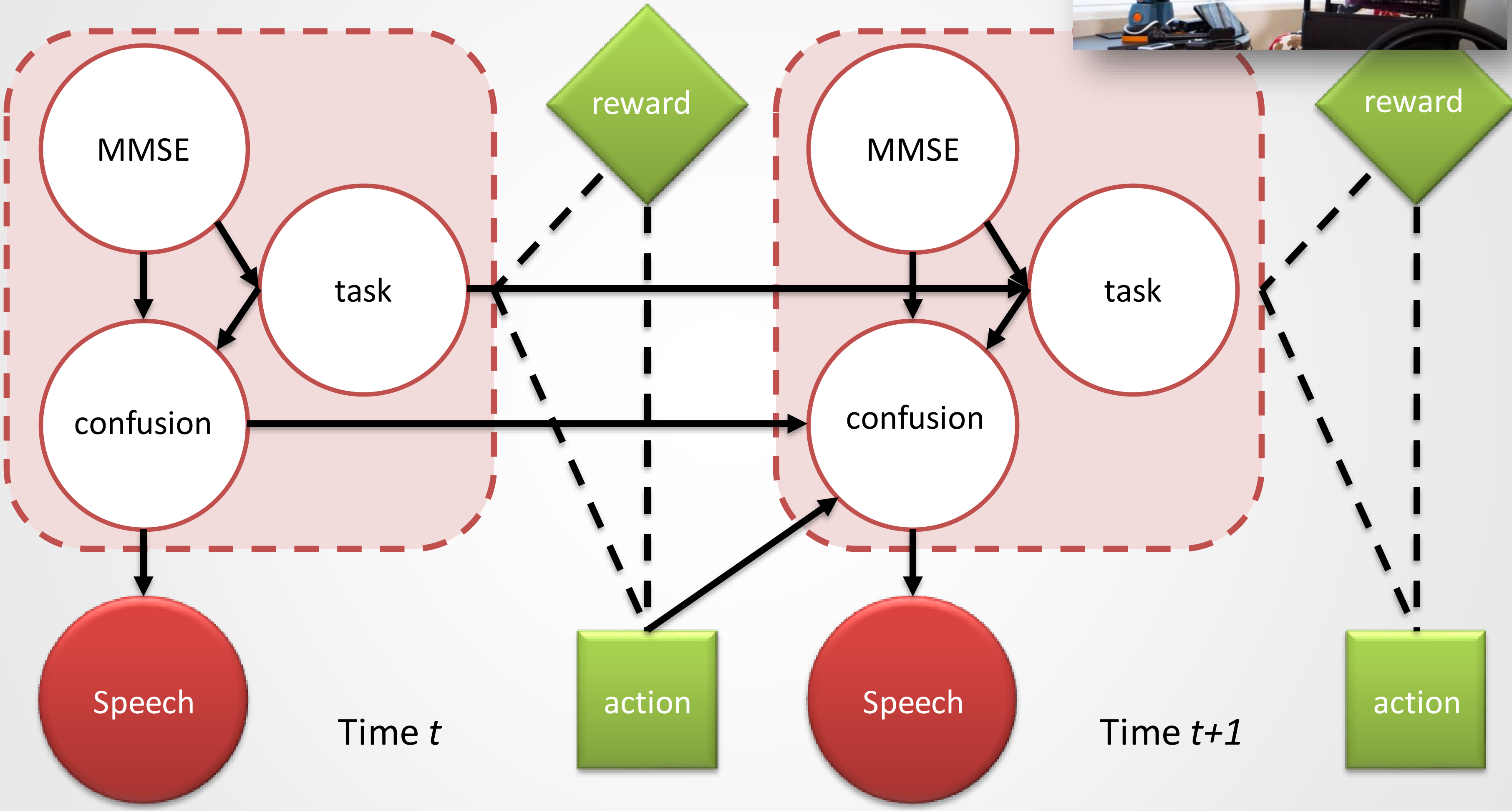
Return $R_t = \sum_{k=t}^{T} \gamma^{(k)} r_k$

Value $V^{\pi}(\mathbf{b}) = \mathrm{E}[R_t | b_t = \mathbf{b}]$

Q $Q^{\pi}(\mathbf{b}, \mathbf{a}) = \mathrm{E}[R_t | b_t = \mathbf{b}, a_t = \mathbf{a}]$

Li J, Monroe W, Ritter A, *et al.* (2017) Deep Reinforcement Learning for Dialogue Generation. doi:10.18653/v1/S17-1008

Aside – RL in dialogue

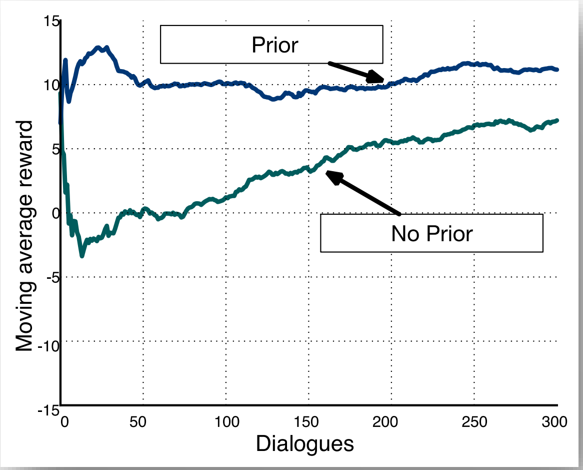


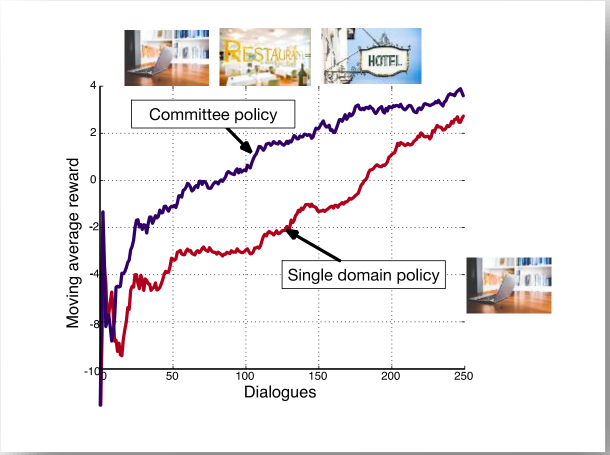
Chinaei H, Currie LC, Danks A, *et al.* (2017) Identifying and avoiding confusion in dialogue with people with Alzheimer's disease. *Computational Linguistics* **43**:377–406.

Aside – RL in dialogue

- Challenge 1 : data is limited in a particular domain
  Solution 1 : learn a distributed architecture with Gaussian priors

- Challenge 2 : Estimates of Q aren't shared across different domains
  Solution 2 : Use a Bayesian 'committee machine'



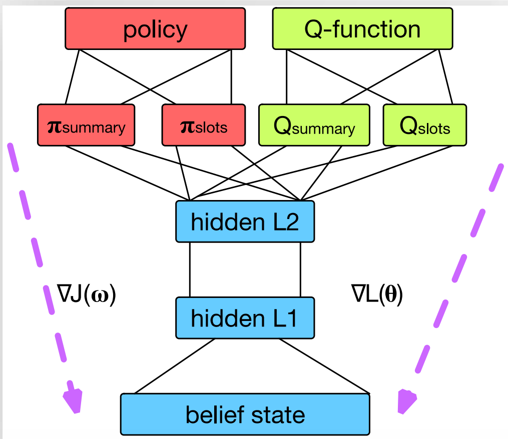


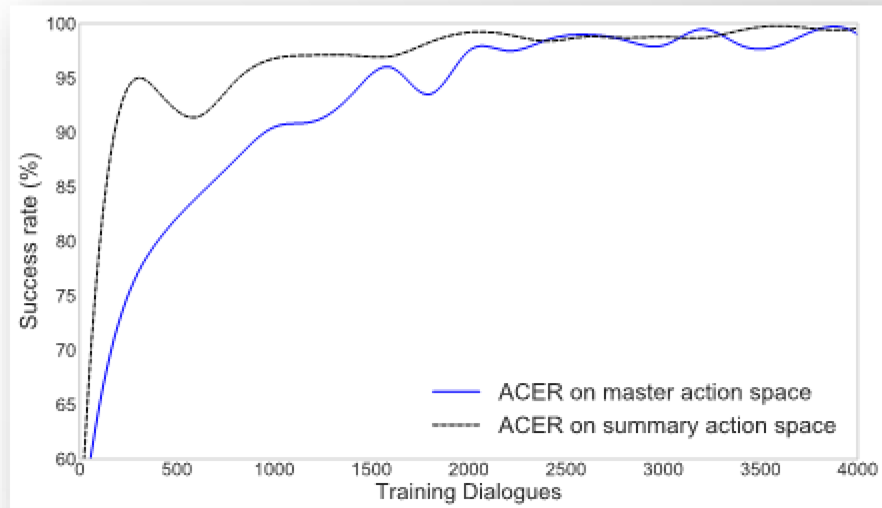
Gašić *et al* (2015) Distributed dialogue policies for multi-domain statistical dialogue management, ICASSP, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7178997
Gašić *et al* (2015) Policy Committee for adaptation in multi-domain spoken dialogue systems, ASRU

UNIVERSITY OF TORONTO

Aside – RL in dialogue

- ACER learns an ‘off policy’ gradient ∇J and modified loss ∇L.
  - Avoid bias through replaying experience





The off-policy version of the Policy Gradient Theorem [30] is used to derive the gradients $\nabla_\omega J(\omega) \approx g(\omega)$:

$$g(\omega) = \sum_{b \in \mathbb{B}} d^\mu(b) \sum_{a \in \mathbb{A}} \nabla_\omega \pi(a|b) Q_\pi(b, a) \tag{1}$$

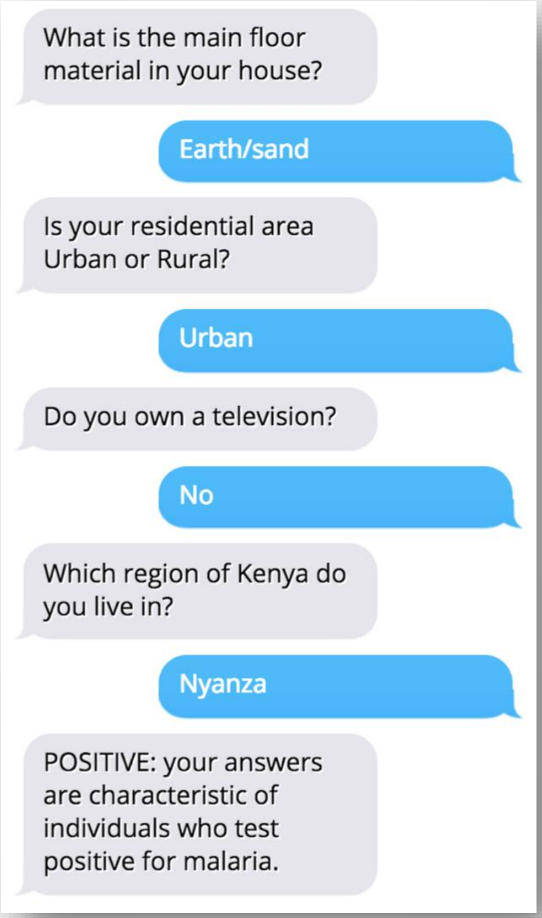
$$\nabla L(\theta) = \nabla_\theta (Q^{ret} - Q_\theta(\mathbf{b}, a))^2$$

$$Q^{ret} = Q(\mathbf{b}, a) + \mathop{\mathbb{E}}_{\mu} \left[\sum_{t \geq 0} \gamma^t \left(\prod_{s=1}^{t} \lambda \min\left(1, \rho(a_s|\mathbf{b}_s)\right) \right) (r_t + \gamma V(\mathbf{b}_{t+1}) - Q(\mathbf{b}_t, a_t)) \right]$$

From Milica Gašić, Cambridge

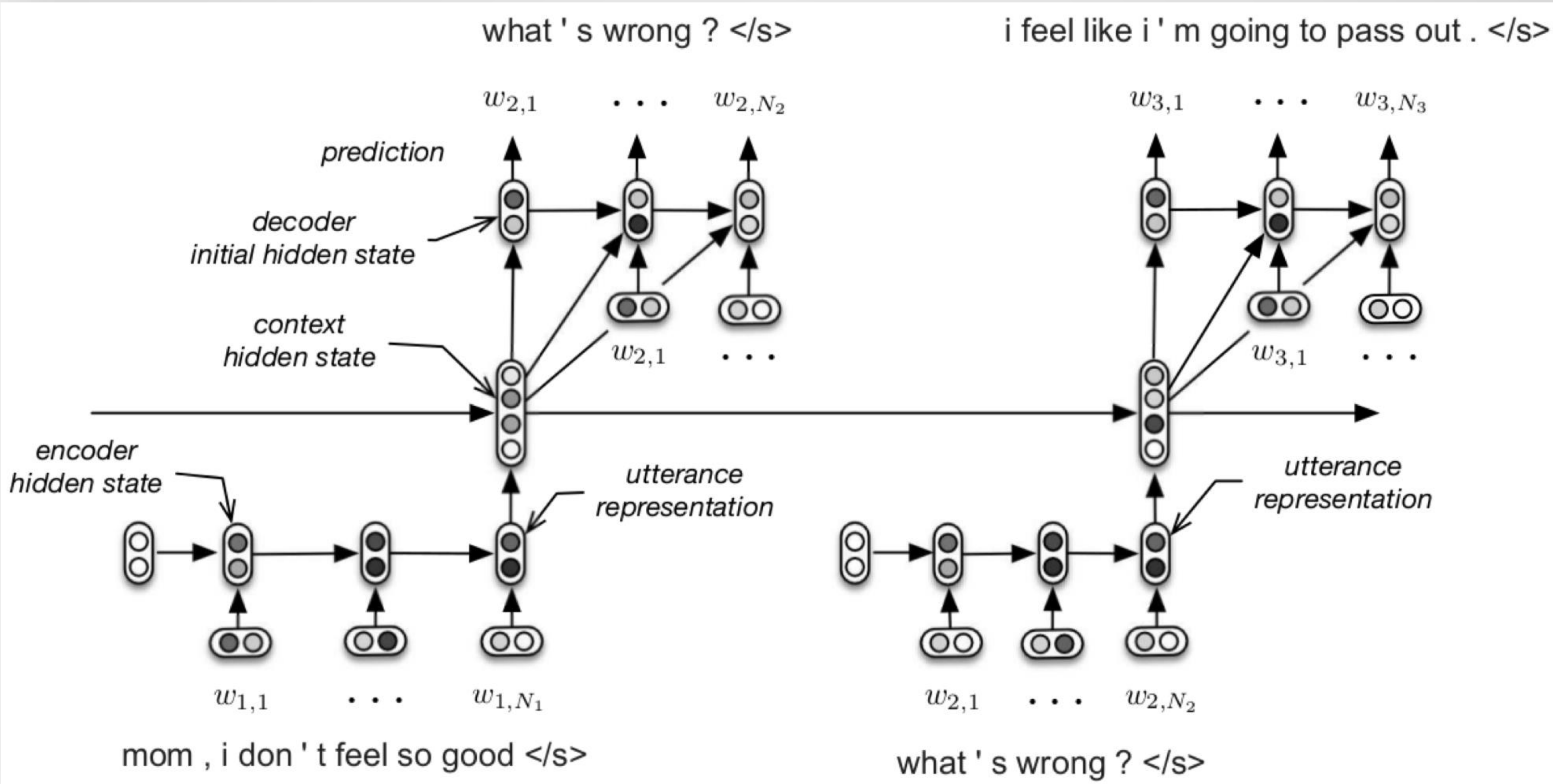
Weisz, Budzianowski, Su, Gašić, (2018) Sample efficient deep reinforcement learning for dialogue systems with large action spaces, IEEE TASLP https://arxiv.org/pdf/1802.03753.pdf

Aside – RL in dialogue



Rajpurkar *et al* (2017) Malaria Likelihood Prediction By Effectively Surveying Households Using Deep Reinforcement Learning. *ML4H*.

End-to-end ~~translation~~ dialogue systems



Serban I V., Sordoni A, Bengio Y, et al. (2015) Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.

Extensions exist that add **variational encoding** or **diversity-promoting objective functions** to avoid Siri-like repetitiveness repetitiveness.

End-to-end dialogue systems

- **Claim**: "we view our model as a **cognitive system**, which has to carry out natural language **understanding, reasoning, decision** making, (*sic*) and natural language generation".
- **Objective**: Perplexity (where U is an utterance)...

$$\exp\left(-\frac{1}{N_w}\sum_{n=1}^{N}\log P_\theta(U_1^n, U_2^n, U_3^n)\right)$$

Serban I V., Sordoni A, Bengio Y, et al. (2015) Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.

- **Overhype** *vb.* make exaggerated claims about (a product, idea, or event) ; publicize or promote excessively

UNIVERSITY OF TORONTO

EVALUATION

Qualitative evaluation

People (sometimes) like cute things that are smaller than they are.

UNIVERSITY OF TORONTO

Corpora for dialogue

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**115,424**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,520,970**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.68
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.18**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**24,071**
# Slots	8	14	4	**61**	13	14	25
# Values	212	1847	99	3871	1363	138	**4510**

Table 1: Comparison of our corpus to similar data sets. Numbers in bold indicate best value for the respective metric. The numbers are provided for the training part of data except for FRAMES data-set were such division was not defined.

- Ubuntu dialogue corpus and AMI Meeting corpus are also popular.

Budzianowski P, Wen T-H, Tseng B-H, *et al.* (2018) MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling http://arxiv.org/abs/1810.00278

Evaluating end-to-end dialogue

- *PyDial* (pydial.org) is an open-source Python toolkit for dialogue evaluation.
  - Domain-independent

- Crowd sourcing (e.g., Mechanical Turk)?
  - Gather many responses to input by humans,
  - Learn to **generate** responses
  - Learn to **discriminate** real from fake.

For number of training iterations **do**
. **For** i=1,D-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Update D using (X, Y) as positive examples and $(X, \hat{Y})$ as negative examples.
. **End**
.
. **For** i=1,G-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Compute Reward r for $(X, \hat{Y})$ using D.
. Update G on $(X, \hat{Y})$ using reward r
. Teacher-Forcing: Update G on (X, Y)
. **End**
End

Figure 1: A brief review of the proposed adversarial reinforcement algorithm for training the generator G and discriminator D. The reward r from the discriminator D can be computed using different strategies according to whether using REINFORCE or REGS. The update of the generator G on $(X, \hat{Y})$ can be done by either using Eq.2 or Eq.3. D-steps is set to 5 and G-steps is set to 1.

Ultes, Rojas-Barahona, Su, *et al* (2017) PyDial: A Multi-domain Statistical Dialogue System Toolkit, ACL, https://www.aclweb.org/anthology/P17-4013
Casanueva, Budzianowski, Su, *et al* (2017) A Benchmarking Environment for Reinforcement Learning Based Task Oriented Dialogue Management, NIPS Symposium on Deep RL
Li J, Monroe W, Shi T, *et al.* (2017) Adversarial Learning for Neural Dialogue Generation. EMNLP, http://arxiv.org/abs/1701.06547

UNIVERSITY OF TORONTO

Evaluating end-to-end dialogue

Input	tell me ... how long have you had this falling sickness ?
Vanilla-SEQ2SEQ	i'm not a doctor.
Adversarial	a few months, i guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-SEQ2SEQ	sammy wrote the test sammy wrote the test .
Adversarial	so he took the pills .

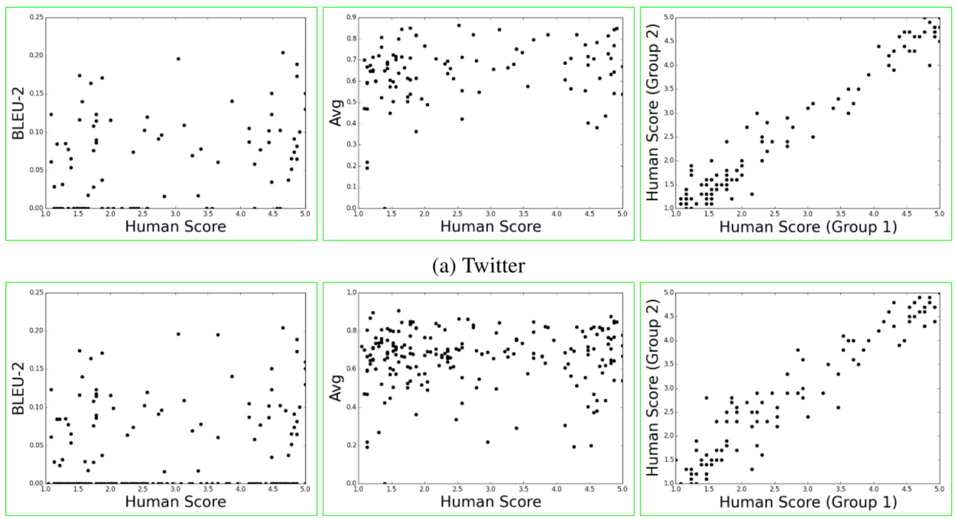
- Evaluating according to scores like BLEU or ROUGE usually require lots of (expensive) **references**.
  - Contribution of **fidelity** can be overwhelmed by **naturalness**.
  - Even still, scores don't correlate *at all* with human judgements.

Li J, Monroe W, Shi T, *et al.* (2017) Adversarial Learning for Neural Dialogue Generation. EMNLP, http://arxiv.org/abs/1701.06547

Evaluating end-to-end dialogue

	Ubuntu Dialogue Corpus			Twitter Corpus		
	Embedding Averaging	Greedy Matching	Vector Extrema	Embedding Averaging	Greedy Matching	Vector Extrema
R-TFIDF	0.536 ± 0.003	0.370 ± 0.002	0.342 ± 0.002	0.483 ± 0.002	0.356 ± 0.001	0.340 ± 0.001
C-TFIDF	0.571 ± 0.003	0.373 ± 0.002	0.353 ± 0.002	0.531 ± 0.002	0.362 ± 0.001	0.353 ± 0.001
DE	$\mathbf{0.650 \pm 0.003}$	0.413 ± 0.002	0.376 ± 0.001	$\mathbf{0.597 \pm 0.002}$	0.384 ± 0.001	0.365 ± 0.001
LSTM	0.130 ± 0.003	0.097 ± 0.003	0.089 ± 0.002	0.593 ± 0.002	$\mathbf{0.439 \pm 0.002}$	$\mathbf{0.420 \pm 0.002}$
HRED	0.580 ± 0.003	$\mathbf{0.418 \pm 0.003}$	$\mathbf{0.384 \pm 0.002}$	$\mathbf{0.599 \pm 0.002}$	$\mathbf{0.439 \pm 0.002}$	$\mathbf{0.422 \pm 0.002}$

Table 2: Models evaluated using the vector-based evaluation metrics, with 95% confidence intervals.



(b) Ubuntu

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

Liu C-W, Lowe R, Serban I V., *et al.* (2016) How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. http://arxiv.org/abs/1603.08023


Goodbye