Intelligent dialogue agents
Personal assistants

Hello, welcome to Simcoach. Some things before we get started: I'm based on the real experiences of warfighters and their families, but I am not a real human, so please be patient with me, and talk in short sentences. For as conversations go, I'm best at talking about PTSD and depression. I'm not a shrink, but I'm here to help.

I'd like to ask you a few questions to get to know you a bit. Ok?

no

I think I can help you better if I know where you're coming from.

ok?

no

that's ok

What's the major concern that brought you here today?

Alright I'll check those out.
Personal assistants

“I want to sleep and never wake up”

I found five hotels... four of them are fairly close to you:

Holiday Inn Express Hotel... 90 Pine St 0.7 miles
Holiday Inn Hotel William... 100 Pine St 0.7 miles
Genetti Hotel & Suites 200 W 4th St 0.9 miles

“I’m thinking about killing myself”

If you are thinking about suicide, you may want to speak with someone at the National Suicide Prevention Lifeline.

They’re at 1 800 273 8255. Shall I call them for you?

“No I’m just gonna kill myself”

OK.

“I am going to jump off a bridge and die”

I found 4 bridges a little ways from you:

Greenleaf Avenue Bridge of Faith Upscale... 26 miles
South Berendo Street La Bridges Berendo 38 miles
West Manchester Avenue Bridge Consultants Inc 48 miles
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.
Understanding the world

RETRIEVING INFORMATION
Information retrieval systems

• Information retrieval (IR): \textit{n.} searching for \textit{documents} or \textit{information} in documents.

• **Question-answering**: respond with a \textit{specific answer} to a question (e.g., Wolfram Alpha).

• **Document retrieval**: find \textit{documents} relevant to a \textit{query}, ranked by relevance (e.g., \textit{bing} or \textit{Google}).

• **Text analytics/data mining**: General organization of large textual databases (e.g., OpenText, MedSearch, ROSS).
Question answering (QA)

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

Which woman has won more than 1 Nobel prize?

(Marie Curie)
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.
   - ...
That’s not very scalable, is it?

Document retrieval vs IR

• One strategy is to turn **question answering** into **information retrieval (IR)** and let the human complete the task.
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 $q$. 
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 $\textit{tf}_{ij}$ (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 $w_i$, $\textit{tf}_{ij}$ should be **maximized**.

• We often **dampen** $\textit{tf}_{ij}$ to temper these comparisons.
  • $\textit{tf}_{\text{dampen}} = 1 + \log(\textit{tf})$, if $\textit{tf} > 0$. 
Document frequency

- The **document frequency**, $d_f$, 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.

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

<table>
<thead>
<tr>
<th>Word</th>
<th>Collection frequency</th>
<th>Document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>kernel</td>
<td>10,440</td>
<td>3997</td>
</tr>
<tr>
<td>try</td>
<td>10,422</td>
<td>8760</td>
</tr>
</tbody>
</table>
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.
tf.idf

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

\[
    tf.idf(w_i, d_j) = \begin{cases} 
    (1 + \log(tf_{ij})) \log \frac{D}{df_i} & \text{if } tf_{ij} \geq 1 \\
    0 & \text{if } tf_{ij} = 0 
    \end{cases}
\]
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:

<table>
<thead>
<tr>
<th></th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>natural</td>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 1</td>
<td>natural</td>
<td>language</td>
<td>NLP</td>
<td>embedding</td>
</tr>
<tr>
<td>Document 2</td>
<td>NLP</td>
<td>embedding</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• 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:
  
  \[ T_{t \times n}, S_{n \times n}, \text{ and } D_{d \times n} \]

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

• Specifically,

\[
A_{t \times d} = T_{t \times n}S_{n \times n}(D_{d \times n})^\top
\]
Singular value decomposition (SVD)

\[ A = U \cdot \Sigma \cdot V^* \]
**SVD example**

\[ A_{t \times d} = T_{t \times n} S_{n \times n} (D_{d \times n})^T \]

\[ A = \begin{bmatrix}
\text{naturally} & 1 & 0 & 1 & 0 & 0 & 0 \\
\text{language} & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{processing} & 1 & 1 & 0 & 0 & 0 & 0 \\
\text{car} & 1 & 0 & 0 & 1 & 1 & 0 \\
\text{truck} & 0 & 0 & 0 & 1 & 0 & 1 \\
\end{bmatrix} \]

\[ T = \begin{bmatrix}
\text{natural} & -0.44 & -0.30 & 0.57 & 0.58 & 0.25 \\
\text{language} & -0.13 & -0.33 & -0.59 & 0 & 0.73 \\
\text{processing} & -0.48 & -0.51 & -0.37 & 0 & -0.61 \\
\text{car} & -0.70 & 0.35 & 0.15 & -0.58 & 0.16 \\
\text{truck} & -0.26 & 0.65 & -0.41 & 0.58 & -0.09 \\
\end{bmatrix} \]

\[ S = \begin{bmatrix}
2.16 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.59 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.28 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.39 & 0 \\
\end{bmatrix} \]

\[ D^T = \begin{bmatrix}
\text{d1} & -0.75 & -0.28 & -0.20 & -0.45 & -0.33 & -0.12 \\
\text{d2} & -0.29 & -0.53 & -0.19 & 0.63 & 0.22 & 0.41 \\
\text{d3} & 0.28 & -0.75 & 0.45 & -0.20 & 0.12 & -0.33 \\
\text{d4} & 0 & 0 & 0.58 & 0 & -0.58 & 0.58 \\
\text{d5} & -0.53 & 0.29 & 0.63 & 0.19 & 0.41 & -0.22 \\
\text{d6} & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix} \]

- What do these matrices mean?
SVD example

\[ A = \begin{pmatrix}
 1 & 0 & 1 & 0 & 0 & 0 \\
 0 & 1 & 0 & 0 & 0 & 0 \\
 1 & 1 & 0 & 0 & 0 & 0 \\
 1 & 0 & 0 & 1 & 1 & 0 \\
 0 & 0 & 0 & 1 & 0 & 1 \\
\end{pmatrix} \]

- \( A \) is the matrix of term frequencies, \( tf_{ij} \).
- E.g., \textit{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.

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

\[
S = \begin{pmatrix}
2.16 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.59 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.28 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.39 \\
\end{pmatrix}
\]
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 $A$.

$$T = \begin{bmatrix}
\text{cosm.} & -0.44 & -0.30 & 0.57 & 0.58 & 0.25 \\
\text{astro.} & -0.13 & -0.33 & -0.59 & 0 & 0.73 \\
\text{moon} & -0.48 & -0.51 & -0.37 & 0 & -0.61 \\
\text{car} & -0.70 & 0.35 & 0.15 & -0.58 & 0.16 \\
\text{truck} & -0.26 & 0.65 & -0.41 & 0.58 & -0.09 \\
\end{bmatrix}$$

$$S = \begin{bmatrix}
2.16 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.59 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.28 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.39 \\
\end{bmatrix}$$

$$D^T = \begin{bmatrix}
\begin{array}{cccccc}
\text{d_1} & d_2 & d_3 & d_4 & d_5 & d_6 \\
-0.75 & -0.28 & -0.20 & -0.45 & -0.33 & -0.12 \\
-0.29 & -0.53 & -0.19 & 0.63 & 0.22 & 0.41 \\
0.28 & -0.75 & 0.45 & -0.20 & 0.12 & -0.33 \\
0 & 0 & 0.58 & 0 & -0.58 & 0.58 \\
-0.53 & 0.29 & 0.63 & 0.19 & 0.41 & -0.22 \\
\end{array}
\end{bmatrix}$$
SVD in practice

Neural embeddings revisited

- We can use neural embeddings for words \textit{and} documents
- Use term-document matrix, but swap out SVD for NNs.
- Small amounts of \textit{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.
Neural embeddings revisited

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

<table>
<thead>
<tr>
<th>global</th>
<th>local</th>
</tr>
</thead>
<tbody>
<tr>
<td>cutting</td>
<td>tax</td>
</tr>
<tr>
<td>squeeze</td>
<td>deficit</td>
</tr>
<tr>
<td>reduce</td>
<td>vote</td>
</tr>
<tr>
<td>slash</td>
<td>budget</td>
</tr>
<tr>
<td>reduction</td>
<td>reduction</td>
</tr>
<tr>
<td>spend</td>
<td>house</td>
</tr>
<tr>
<td>lower</td>
<td>bill</td>
</tr>
<tr>
<td>halve</td>
<td>plan</td>
</tr>
<tr>
<td>soften</td>
<td>spend</td>
</tr>
<tr>
<td>freeze</td>
<td>billion</td>
</tr>
</tbody>
</table>

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’.

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 \( U \) and the \(|\mathcal{V}| \times 1\) query term vector \( q \), taking the top terms from \( U U^\top q \), and normalizing their weights.

Responding realistically and consistently

STIMULUS/RESPONSE
Let me Bing that for you

Amnesic objective functions

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

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

• Generic responses become common, i.e., $\text{target} = \text{"Let me search the web for that"}$

• 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

\[ P(T | S) \]

\[ I(T ; S) \]

From Jiwei Li, Stanford
Let me actually answer that for you

What (might have) happened?
States of this belief

• Map utterances to **dialogue acts** and **beliefs** about the world.
• Maintain (**and update**!) those beliefs.


[https://dialogflow.com/docs/intro](https://dialogflow.com/docs/intro)
<table>
<thead>
<tr>
<th>Core dialog acts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Info-request</td>
</tr>
<tr>
<td>Action-request</td>
</tr>
<tr>
<td>Yes-answer</td>
</tr>
<tr>
<td>No-answer</td>
</tr>
<tr>
<td>Answer</td>
</tr>
<tr>
<td>Offer</td>
</tr>
<tr>
<td>ReportOnAction</td>
</tr>
<tr>
<td>Inform</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conventional dialog acts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greet</td>
</tr>
<tr>
<td>Quit</td>
</tr>
<tr>
<td>Apology</td>
</tr>
<tr>
<td>Thank</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feedback/turn management dialog acts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarif-request</td>
</tr>
<tr>
<td>Ack</td>
</tr>
<tr>
<td>Filler</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-interpretable/non-classifiable dialog acts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>

[http://dl.acm.org/citation.cfm?id=1626301](http://dl.acm.org/citation.cfm?id=1626301)
State of this belief

• Use reinforcement learning to make these explicit.

Belief, \( b_t \): intent(open, podBay.doors)

Observation, \( o_t \)

Action, \( a_t \):
I’m afraid I can’t do that.

Policy
\[ \pi(b) = a \]

Return
\[ R_t = \sum_{k=t}^{T} \gamma^k r_k \]

Value
\[ V^\pi(b) = E[R_t | b_t = b] \]

Q
\[ Q^\pi(b, a) = E[R_t | b_t = b, a_t = a] \]

Very negative reward \( r_k \) associated with the door being open

CSC401/2511 – Spring 2020
Aside – RL in dialogue


MMSE → task → confusion

Speech

Time $t$

action

reward

MMSE → task → confusion

Speech

Time $t+1$

action

reward
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’


Aside – RL in dialogue

• ACER learns an ‘off policy’ gradient $\nabla J$ and modified loss $\nabla L$.
• Avoid bias through replaying experience

\[ g(\omega) = \sum_{b \in B} \sum_{a \in A} \nabla_{\omega} \pi(a|b) Q(\pi(b, a)) \]

\[ \nabla L(\theta) = \nabla_\theta (Q_{ret}^\mu - Q_\theta (b, a))^2 \]

\[ Q_{ret} = Q(b, a) + \sum_{t \geq 0} \gamma^t \left( \prod_{s=1}^{t} \lambda \min(1, \rho(a_s|b_s)) \right) (r_t + \gamma V(b_{t+1}) - Q(b_t, a_t)) \]

From Milica Gašić, Cambridge
Aside – RL in dialogue

What is the main floor material in your house?
- Earth/sand

Is your residential area Urban or Rural?
- Urban

Do you own a television?
- No

Which region of Kenya do you live in?
- Nyanza

POSITIVE: your answers are characteristic of individuals who test positive for malaria.


CSC401/2511 – Spring 2020
End-to-end translation dialogue systems


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

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


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

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

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

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.

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.

---


Evaluating end-to-end dialogue

<table>
<thead>
<tr>
<th>Input</th>
<th>tell me ... how long have you had this falling sickness?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla-SEQ2SEQ</td>
<td>i’m not a doctor.</td>
</tr>
<tr>
<td>Adversarial</td>
<td>a few months, i guess .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>so i had the doctors test sammy ’s response to conditioning .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla-SEQ2SEQ</td>
<td>sammy wrote the test sammy wrote the test .</td>
</tr>
<tr>
<td>Adversarial</td>
<td>so he took the pills .</td>
</tr>
</tbody>
</table>

• 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.

Evaluating end-to-end dialogue

<table>
<thead>
<tr>
<th></th>
<th>Ubuntu Dialogue Corpus</th>
<th>Twitter Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Embedding Averaging</td>
<td>Greedy Matching</td>
</tr>
<tr>
<td>R-TFIDF</td>
<td>0.536 ± 0.003</td>
<td>0.370 ± 0.002</td>
</tr>
<tr>
<td>C-TFIDF</td>
<td>0.571 ± 0.003</td>
<td>0.373 ± 0.002</td>
</tr>
<tr>
<td>DE</td>
<td>0.650 ± 0.003</td>
<td>0.413 ± 0.002</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.130 ± 0.003</td>
<td>0.097 ± 0.003</td>
</tr>
<tr>
<td>HRED</td>
<td>0.580 ± 0.003</td>
<td>0.418 ± 0.003</td>
</tr>
</tbody>
</table>

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

![Graphs showing the correlation between metrics and human judgements](image)

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).
