Intelligent dialogue agents
Personal assistants

![SimCoach interface](image)

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. Far 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?

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

Ok?

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 Hot...
- Holiday Inn Hotel William...
- Genetti Hotel & Suites

"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...
- South Berendo Street La Bridges Berendo
- West Manchester Avenue Bridge Consultants Inc

26 miles
38 miles
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): \( n \) 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)

Which woman has won more than 1 Nobel prize?

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

(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.
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 $tf_{ij}$ (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.
  • $tf_{\text{dampen}} = 1 + \log(tf)$, 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.

<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>

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

\[
\text{tf.idf}(w_i, d_j) = \begin{cases} 
(1 + \log(tf_{ij})) \log \frac{D}{df_i} & \text{if } tf_{ij} \geq 1 \\
0 & \text{if } tf_{ij} = 0
\end{cases}
\]
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><em>Query</em></td>
<td>natural</td>
<td>language</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Document 1</em></td>
<td>natural</td>
<td>language</td>
<td>NLP</td>
<td>embedding</td>
</tr>
<tr>
<td><em>Document 2</em></td>
<td></td>
<td></td>
<td>NLP</td>
<td>embedding</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

The equation for the SVD factorization is:

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

#### Matrices

**$A$**

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>language</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>processing</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>truck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**$T$**

<table>
<thead>
<tr>
<th></th>
<th>nat.</th>
<th>lang.</th>
<th>proc.</th>
<th>car</th>
<th>truck</th>
</tr>
</thead>
<tbody>
<tr>
<td>nat.</td>
<td>-0.44</td>
<td>-0.30</td>
<td>0.57</td>
<td>0.58</td>
<td>0.25</td>
</tr>
<tr>
<td>lang.</td>
<td>-0.13</td>
<td>-0.33</td>
<td>-0.59</td>
<td>0</td>
<td>0.73</td>
</tr>
<tr>
<td>proc.</td>
<td>-0.48</td>
<td>-0.51</td>
<td>-0.37</td>
<td>0</td>
<td>-0.61</td>
</tr>
<tr>
<td>car</td>
<td>-0.70</td>
<td>0.35</td>
<td>0.15</td>
<td>-0.58</td>
<td>0.16</td>
</tr>
<tr>
<td>truck</td>
<td>-0.26</td>
<td>0.65</td>
<td>-0.41</td>
<td>0.58</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

**$S$**

<table>
<thead>
<tr>
<th></th>
<th>2.16</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.59</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1.28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.39</td>
<td>0</td>
</tr>
</tbody>
</table>

**$D^\top$**

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.75</td>
<td>-0.28</td>
<td>-0.20</td>
<td>-0.45</td>
<td>-0.33</td>
<td>-0.12</td>
<td></td>
</tr>
<tr>
<td>-0.29</td>
<td>-0.53</td>
<td>-0.19</td>
<td>0.63</td>
<td>0.22</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>0.28</td>
<td>-0.75</td>
<td>0.45</td>
<td>-0.20</td>
<td>0.12</td>
<td>-0.33</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.58</td>
<td>0</td>
<td>-0.58</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>-0.53</td>
<td>0.29</td>
<td>0.63</td>
<td>0.19</td>
<td>0.41</td>
<td>-0.22</td>
<td></td>
</tr>
</tbody>
</table>

• What do these matrices mean?
**SVD example**

\[ A = \]

<table>
<thead>
<tr>
<th></th>
<th>(d_1)</th>
<th>(d_2)</th>
<th>(d_3)</th>
<th>(d_4)</th>
<th>(d_5)</th>
<th>(d_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>language</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>processing</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>truck</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- \(A\) 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.

$T = \begin{pmatrix}
-0.44 & -0.30 & 0.57 & 0.58 & 0.25 \\
-0.13 & -0.33 & -0.59 & 0 & 0.73 \\
-0.48 & -0.51 & -0.37 & 0 & -0.61 \\
-0.70 & 0.35 & 0.15 & -0.58 & 0.16 \\
-0.26 & 0.65 & -0.41 & 0.58 & -0.09 \\
\end{pmatrix}$

$D^\top = \begin{pmatrix}
-0.75 & -0.28 & -0.20 & -0.45 & -0.33 & -0.12 \\
-0.29 & -0.53 & -0.19 & 0.63 & 0.22 & 0.41 \\
0.28 & -0.75 & 0.45 & -0.20 & 0.12 & -0.33 \\
0 & 0 & 0.58 & 0 & -0.58 & 0.58 \\
-0.53 & 0.29 & 0.63 & 0.19 & 0.41 & -0.22 \\
\end{pmatrix}$

- $T$ 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.

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

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

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

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

*Communications of the ACM* 8:627-633.
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.

Neural embeddings revisited

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

Aside – query expansion

- **Query expansion** involves reweighting likelihoods, usually through **deleted interpolation**:
  \[ p^1_q(w) = \lambda p(w) + (1 - \lambda) p^+_q(w) \]

- \( P^+_q \) 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 \( UU^T q \), and normalizing their weights.

---

Responding realistically and consistently

STIMULUS/RESPONSE
Let me Bing that for you

(a) [Image of phone screen with text: "Find restaurants near me"]

(b) [Image of phone screen with text: "Tell me more about the second one"]

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., 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

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

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
[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.


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

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

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

Q \( Q^\pi(b, a) = \mathbb{E}[R_t | b_t = b, a_t = a] \)
Aside – RL in dialogue

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

From Milica Gašić, Cambridge
Aside – RL in dialogue

End-to-end translation dialogue systems

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)...

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


- **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>14.60</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>13.18</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>61</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>4510</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

**Input**
- tell me ... how long have you had this falling sickness?
- i’m not a doctor.
- a few months, i guess.

**Input**
- so i had the doctors test sammy’s response to conditioning.
- sammy wrote the test sammy wrote the test.
- 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.

Evaluating end-to-end dialogue
