Conversational Scaffolding: An Analogy-based Approach to Response Prioritization in Open-domain Dialogs

Will Myers[1], Tyler Etchart[1] and Nancy Fulda[2]
[1]*Perception, Control and Cognition Laboratory, Brigham Young University, 3361 TMCB, Provo, Utah, U.S.A.*
[2]*DRAGN Labs, Brigham Young University, 3361 TMCB, Provo, Utah, U.S.A.*

Keywords: Response Prioritization, Utterance Retrieval, Dialog Modeling, Natural Language Processing, Vector Space Models, Word Embeddings, Conversational AI, Sentence Representations, Analogical Reasoning.

Abstract: We present *Conversational Scaffolding*, a response-prioritization technique that capitalizes on the structural properties of existing linguistic embedding spaces. Vector offset operations within the embedding space are used to identify an ‘ideal’ response for each set of inputs. Candidate utterances are scored based on their cosine distance from this ideal response, and the top-scoring candidate is selected as conversational output. We apply our method in an open-domain dialog setting and show that the most effective analogy-based strategy outperforms both an Approximate Nearest-Neighbor approach and a naive nearest neighbor baseline. We also demonstrate the method’s ability to retrieve relevant dialog responses from a repository containing 19,665 random sentences. As an additional contribution we present the Chit-Chat dataset, a high-quality conversational dataset containing 483,112 lines of friendly, respectful chat exchanges between university students.

1 INTRODUCTION

High-quality conversational data is a scarce resource even in the modern internet era. Unmoderated online interactions, while plentiful and easy to harvest, fail to exhibit the verbal patterns, topical continuity, and social restraint that one would like to see in a personal assistant or other conversational agent. The prevalence of trolls in online chat forums and social media platforms is particularly problematic (Buckels et al., 2014) (Rainie and Anderson, 2017), especially when anonymous or pseudonymous commenting is possible (Cho and Acquisti, 2013). Even in heavily moderated contexts or in datasets that have been hand-curated to filter out derogatory posts, the length and content of comments may differ drastically from natural conversation (Schneider et al., 2002). Data harvested from recordings or phone conversations is subject to transcription errors, rambling thoughts, and incomplete sentences, while dialogs extracted from novels or movie scripts run the risk of being melodramatic and inauthentic. (It would seem odd indeed if a personal assistant trained using such data were to profess undying love toward its conversation partner.)

Given this data scarcity, researchers face the question: How shall automated systems mimic human behavior when so little source information is available?

We address this question by first observing that, while language is combinatorial in nature and thus able to represent a nearly infinite span of ideas, the *patterns* of language are far more tractable. Certain types of statements encourage certain types of responses, regardless of the specific conversation topic. These patterns can be detected and imitated via the use of analogical relations within a pre-trained embedding space. Thus, a relatively small corpus of exemplars can be used to guide the response ranking system of a conversational agent.

The basic concept is simple: We begin by encoding a reference corpus, called our *scaffold*, using one of many available pre-trained embedding models. Incoming utterances are matched against the scaffold corpus based on the embedded concatenation of the utterances in the dialog history, and the top n contextual matches are used to calculate an analogically coherent response, or *target point* within the embedding space. The candidate utterance with the lowest cosine distance from the target point is selected as the agent’s dialog response.

We examine the effectiveness of our algorithm on a response prediction task and show that the most effective vector offset method outperforms both an Approximate Nearest Neighbor classifier and a naive nearest-neighbor approach. We then apply our scaf-

Myers, W., Etchart, T. and Fulda, N.
Conversational Scaffolding: An Analogy-based Approach to Response Prioritization in Open-domain Dialogs.
DOI: 10.5220/0008939900690078
In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020) - Volume 2*, pages 69-78
ISBN: 978-989-758-395-7; ISSN: 2184-433X

folding technique to a real-time conversational scenario using a new, high-quality conversational corpus called the Chit-Chat dataset. We show that the resulting automated responses, while not perfect, nevertheless mimic the style and flow of human conversation.

2 RELATED WORK

Retrieval systems for conversational AI have historically relied on statistical models such as Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and McGill, 1986) (Gandhe and Traum, 2013) (Charras et al., 2016) or token-level vector space models (Banchs and Li, 2012) (Dubuisson Duplessis et al., 2017) (Charras et al., 2016), with cosine distance used to score candidate utterances. More recently, a variety of neural models for information retrieval have been explored, including a paraphrase matching algorithm utilizing recursive auto-encoders (Nio et al., 2016), models based on LSTMs and other recurrent units (Lowe et al., 2017) (Nio et al., 2016), and sequential matching networks that use an RNN to accumulate vectors representing the relationship between each response and the utterances in the context (Wu et al., 2017). The potential effectiveness of such methods is limited, however, by the relative scarcity of high-quality conversational training data.

2.1 Semantic Comparisons

The recent availability of general purpose pre-trained linguistic embedding spaces opens new possibilities for utterance retrieval. Sentence-level embedding spaces such as skip-thought vectors (Kiros et al., 2015), quick-thought vectors (Logeswaran and Lee, 2018), InferSent (Conneau et al., 2017), and Google's Universal Sentence Encoder (Cer et al., 2018), as well as task-specific embeddings extracted from popular generative models such as HRED (Serban et al., 2016) and GPT-2 (Radford et al.,), are often used to approximate the semantic distance between two sentences. For example, Bartl et al. use conversational contexts extracted from HRED embeddings to retrieve a set of *n* candidate sentences via an Approximate Nearest Neighbor (ANN) algorithm (Bartl and Spanakis, 2017). Each candidate is then scored based on its cosine similarity to each of the other candidates in the retrieved set.

Alas, linguistic embedding models do not strictly encode semantic structure (Thalenberg, 2016) (Dou et al., 2018) (Kim et al., 2016), and the use of cosine distance as an approximation of semantic similarity is only partially successful (Fu et al., 2018) (Patro et al., 2018). The development of embedding spaces with improved semantic structure is an active area of research (Zhu et al., 2018) (Conneau et al., 2018), however, we circumvent this limitation by relying on the cosine distances between vector offsets, rather than between the embeddings themselves, in order to calculate our target point.

2.2 The Analogical Structure of Embedding Spaces

Our dialog response ranking algorithm leverages the analogical structure inherent in language, and by extension also inherent in linguistic embedding spaces, to improve response selection. In many computer science communities it is commonly known that word embeddings such as word2vec (Mikolov et al., 2013a), GLoVE (Pennington et al., 2014), and FastText (Bojanowski et al., 2016) can be used to solve linguistic analogies of the form a:b::c:d. This is generally accomplished using vector offsets such as [*Madrid* - *Spain* + *France* ≈ *Paris*] or [*walking* - *walked* + *swimming* ≈ *swam*] (Mikolov et al., 2013b) (Gladkova et al., 2016). Query accuracy can be further improved by averaging multiple vector offsets (Drozd et al., 2016) (Fulda et al., 2017a) or by extending the length of the offset vector (Fulda et al., 2017b).

Our research extends this notion of analogical relationships into the realm of multi-word embeddings. We postulate (and show via our results) that sentence-level embedding spaces can contain similar analogical relationships, and that these relationships can be utilized to select plausible responses in open-domain dialogs. Thus, rather than evaluating candidate responses based on their strict distance to exemplars in the scaffold corpus, we instead rely on the relative distance between pairs of sentences in order to locate an idealized response vector which corresponds to point d in the classic a:b::c:d analogical form. Candidate responses are scored based on their cosine distance from this target point.

3 ALGORITHM

Our conversational scaffolding algorithm, first introduced as an experimental sub-component of (Fulda et al., 2018), is expanded and refined in this work by clarifying the localization technique and by performing a structured evaluation of response accuracy across a variety of scoring algorithms, including new baselines. We also provide an example of the algorithm in action, showing its ability to retrieve relevant

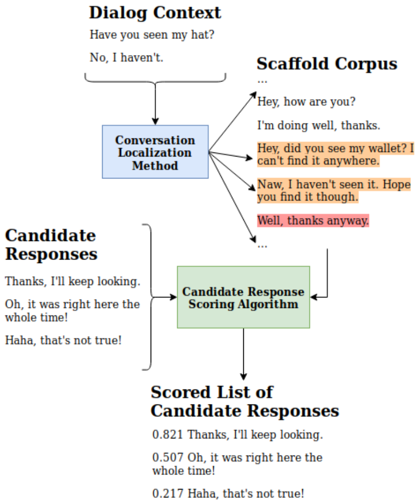
Dialog Context
Have you seen my hat?
No, I haven't.
Conversation Localization Method
Scaffold Corpus
...
Hey, how are you?
I'm doing well, thanks.
Hey, did you see my wallet? I can't find it anywhere.
Naw, I haven't seen it. Hope you find it though.
Well, thanks anyway.
...
Candidate Responses
Thanks, I'll keep looking.
Oh, it was right here the whole time!
Haha, that's not true!
Candidate Response Scoring Algorithm
Scored List of Candidate Responses
0.821 Thanks, I'll keep looking.
0.507 Oh, it was right here the whole time!
0.217 Haha, that's not true!

Figure 1: Workflow diagram: The dialog context is converted to an array of sentence embeddings using Google's Universal Sentence Encoder, then passed to an embedded concatenation localization function to determine the best contextual match(es). The matched utterances (orange) along with their direct successors (red) are then passed to the Response Scoring Algorithm, which assigns a numerical value to each candidate response.

responses from a set of ca. 20,000 randomly-selected sentences.

Figure 1 gives an overview of our methodology. Given a dialog context of variable length, our algorithm first locates a set of high-quality contextual matches within the scaffold corpus. These contextual matches, along with the utterance directly following each context match, are then passed to one of several scoring algorithms. All utterances are encoded using Google's Universal Sentence Encoder Lite (USE Lite) (Cer et al., 2018), a lightweight but impressively robust embedding model. We selected this model based on its unusually high performance as a heuristic for semantic distance. Preliminary experiments in our laboratory revealed that USE Lite was able to achieve a Pearson's r score of 0.752 on the 2013 Semantic Textual Similarity benchmark, the highest score of any model we tried. Other model scores were: InferSent (0.718), FastText bag-of-words (0.547), BERT (0.495), GloVe bag-of-words (0.404), skip-thought vectors (0.214), and GPT-2 (-0.052).

3.1 Contextual Alignment

Contextual alignment refers to the process of matching incoming utterances against similar utterance patterns within a scaffold corpus. This can be done naively by using an Approximate Nearest Neighbor[1] algorithm based on a simple Euclidean distance metric[2]. In this paradigm, for a dialog history of length n, the optimal contextual match can be identified using the expression

$$min_z \sum_{i=1}^{n} ||v_i - s_{z+i}|| \tag{1}$$

where $\{v_1, \ldots, v_n\}$ are the vector embeddings of the n most recent utterances in the current dialog and $\{s_{z+1}, \ldots, s_{z+n}\}$ represent the vectors located within a sliding window of length n beginning at element z of the pre-embedded scaffold corpus. The notation $||x||$ represents the Euclidean norm of vector x.

This Euclidean distance approach is easy to calculate, but it ignores the powerful analogical structure inherent within the embedding space. For example, assume we have the following dialog history coupled with two potential contextual matches in the scaffold corpus:

dialog history

```
1.  Did you watch the basketball game?
2.  Yeah, that slam dunk at the end was really
impressive.
```

contextual match A

```
1.  Did you watch the football game?
2.  Yeah, that touchdown at the end was really
impressive.
```

contextual match B

```
1.  Did you watch the basketball game?
2.  Yeah, that was an impressive game.
```

Using the example text, a Euclidean distance approach using Google's Universal Sentence Encoder Lite and Eq. 1 above will select contextual match B (with a summed distance of 0.884) over A (which has a summed distance of 1.191). And yet in many ways, contextual match A is a closer semantic parallel to the actual dialog history. In particular, the conversational pattern exhibited in A is an almost perfect match for the dialog, even though the specific topic differs.

In order to capture such subtleties, we propose an alternate method of contextual alignment: *Embedded Concatenation*. Embedded Concatenation leverages the structure of the embedding space by concatenating the input sentences prior to encoding them via Universal Sentence Encoder Lite (Cer et al., 2018). A naive

[1] Approximate approaches, rather than a more rigorous K-Nearest Neighbor algorithm, are used in order to improve computation speed.

[2] Any valid distance metric, such as cosine distance, can be used. We tested both cosine and Euclidean distance, but found the latter to be empirically better.

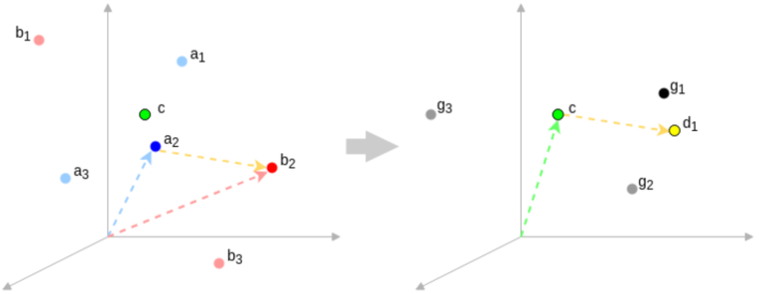

Figure 2: Naive Analogy, where: c (green) represents the embedded input utterance, a_i (blue) represent the nearest embedded utterances from the scaffold corpus, b_i (red) represent the embedded successors to a_i in the scaffold corpus, $d_1 = c + b_2 - a_2$ (yellow) represents the 'ideal' response, and g_i (grey and black) represent embedded candidate responses with g_1 (black) representing the response selected by the naive-analogy scoring algorithm. Image originally from (Fulda et al., 2018).

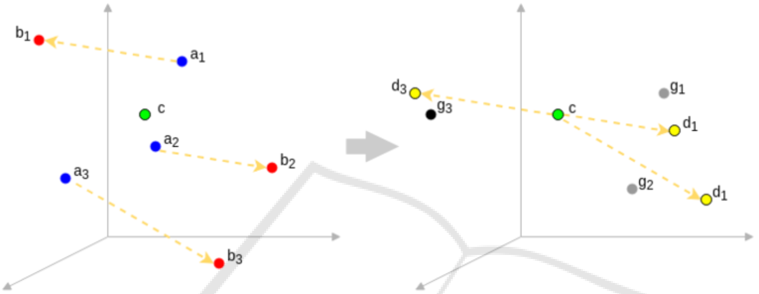

Figure 3: Scattershot Method, where: c (green) represents the embedded input utterance, a_i (blue) represent the nearest embedded utterances from the scaffold corpus, b_i (red) represent the associated embedded successors to a_i in the scaffold, $d_i = c + (b_i - a_i)$ (yellow) represent the 'ideal' responses, and g_i (grey and black) represent embedded candidate responses with g_3 (black) representing the response selected by the scattershot scoring algorithm. Image originally from (Fulda et al., 2018).

Euclidean distance metric is then used to match the embedded concatenation against each element in the pre-embedded scaffold corpus. The optimal contextual match is:

$$min_z \ ||embed(h_1 + ... + h_n) - s_z|| \qquad (2)$$

where $\{h_1, ..., h_z\}$ are the plain text (i.e. *un*embedded) utterances in the dialog history, the $+$ symbol represents string concatenation (with an extra space inserted between sentences), s_z is an arbitrary vector located within the pre-embedded scaffold corpus, and $embed(x)$ denotes the process of embedding a plain text utterance x to obtain its corresponding vector representation.

Note that the described localization method assumes that only a single, optimal, contextual match is desired. This was done for simplicity. In reality, it is often beneficial to take the k best matches, and in fact many of the scoring algorithms in section 3.2 require $k > 1$. The diagrams in Section 3.2 assume a value of $k = 3$ for clarity. In our empirical experiments, a value of $k = 5$ was used.

3.2 Candidate Response Scoring

Once the top k contextual matches for the dialog history have been identified, the candidate responses can be scored. Candidate responses may come from a repository of pre-selected utterances, or they may be produced dynamically via generative models, scripted templates, or other text generation methods. For ease of representation, the algorithm descriptions in Figures 2-4 assume a conversation history of length 1 combined with a Euclidean distance approach to conversational localization. Extensions to longer conversation histories and to the embedded concatenation localization method are straightforward and easy to implement.

In Figs 2-4, the use of the letters a_i, b_i, c, and d_i corresponds to the classic linguistic analogy structure a:b::c:d ('a is to b as c is to d'), which can be solved using vector offsets of the form $c + b - a = d$.

1. *Naive Analogy* (Fig 2). This scoring algorithm represents the simplest possible use of analogical structure when scoring candidate responses, and is an extension at the sentence level of the classic a:b::c:d analogies used in conjunction with word em-

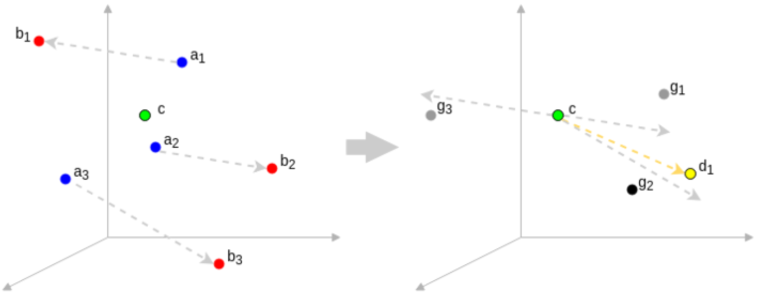

Figure 4: Flow Vectors Method, where: c (green) represents the embedded input utterance, a_i (blue) represent the nearest embedded utterances from the scaffold corpus, b_i (red) represent the associated embedded successors to a_i in the scaffold corpus, $d_1 = c + 1/n\sum(b_i - a_i)$ (yellow) represents the 'ideal' response, and g_i (grey and black) represent embedded candidate responses with g_2 (black) representing the response selected by the flow vectors scoring algorithm. Image originally from (Fulda et al., 2018).

beddings (Mikolov et al., 2013b) (Gladkova et al., 2016). Using a value of $k = 1$, the naive analogy locates the single best context match within the scaffold corpus, along with its (pre-embedded) successor. The vector difference between the successor and the last sentence in the context window is then added to the embedded vector representation of the most recent utterance in the dialog history. Candidate utterances are scored based on their distance from the resulting point in vector space.

2. *Scattershot* (Fig 3). The scattershot scoring algorithm takes the non-deterministic nature of language into account by assuming that there are many valid responses. It therefore searches for a candidate response that matches *any* of several high-scoring context matches. In this method, the vector differences between each context match and its respective successor are calculated separately, and then added to the vector embedding of the most recent utterance in the dialog history. The result is a set of k target points, each of which represents a possible valid conversational response. The candidate located nearest to *any one* of these points receives the highest score.

3. *Flow Vectors* (Fig 4). Lastly, the flow vectors algorithm presumes that there is some manifold of acceptable responses within the embedding space, and seeks to calculate the centroid of that manifold by averaging the differences between multiple context matches and their successors. The candidate response nearest to this averaged centroid receives the highest score.

4 THE CHIT-CHAT DATASET

To be effective, our algorithmic approach requires a high-quality conversational dataset to be used as a scaffold. The scaffold dataset will define the style and feel of the human-computer interaction, so its contents should mimic genuine, in-person conversations as much as possible. It should also cover a wide range of topics, be free from offensive or trollish behavior, include fluid and natural-sounding sentences, and not be unduly cluttered with web-links, marketing material, or emoticons. Unsatisfied with the dataset options available at the time we commenced our research, we elected to construct our own.

The **Chit-Chat dataset** (https://github.com/BYU-PCCL/chitchat-dataset) is a novel dataset generated using a university-wide conversational competition launched in 2018. Working in conjunction with a marketing team, we created a website that would randomly pair students together and let them chat via a simple text-message interface. Participants' chats were scored based on the number of words used per turn, number of turns, absence of incendiary language, use of correct grammar, and so forth. The highest scoring chatters received prizes.

The resulting dataset contains 482,112 dialog turns from over 1200 users, with a vocabulary size of 85,952. A key element of the Chit-Chat dataset is the commitment, made by each participant when signing up for the competition, that they would abstain from mean-spirited or offensive behaviors. While this is difficult to police algorithmically, post-competition examinations of the chat data reveal that by and large, the students held to their agreement. They tried to seek common ground rather than points of discord, and when they disagreed with each other, they did so in a respectful and non-inflammatory manner. We have found the Chit-Chat conversations highly useful, and are releasing the dataset publicly in hopes that it will help other researchers.

5 EXPERIMENTS

In current state-of-the-art conversational systems, candidate utterances may be obtained via generative models (Sutskever et al., 2014) (Vaswani et al., 2017), template-based algorithms (Fulda et al., 2018), or pre-built language repositories (Wu et al., 2017) (Al-Zubaide and Issa, 2011). Our experimental setup is based on the latter case, although the same algorithms could easily be used to prioritize over utterances produced by hand-coded scripts or by an ensemble of neural generators.

We structure our empirical evaluation as a classification task with six candidate responses. One of the responses is the actual next sentence in the dialog. The other five candidates are randomly-chosen distractors drawn from the same dialog corpus as the conversation history. The task of the agent is to leverage the dialog patterns in the scaffold corpus to identify the correct response. All experiments in this section used a context length $n = 2$ and considered the $k = 5$ best matches for each scaffolding algorithm.

5.1 Text Corpora

To simulate an open-domain dialog setting, we needed candidate utterances with a broad spread of conversation topics and dialog styles. To achieve this, we merged data from four different sources:

1. Chit-Chat[3]
2. Daily Dialog[4] (Li et al., 2017)
3. A 33 million word subset of Reddit[5]
4. Ubuntu Dialogue Corpus[6] (Lowe et al., 2015)

The Chit-Chat dataset, collected locally via a university competition, contains 483,112 dialog turns between university students using an informal online chat framework. The Daily Dialog dataset simulates common, real-life interactions such as shopping or ordering food at a restaurant. Reddit[7] covers an array of general topics, with copious instances of web links, internet acronyms, and active debate. Finally, the Ubuntu Dialogue Corpus contains 966,400 dialog turns taken from the Ubuntu Chat Logs, with a heavy emphasis on troubleshooting and technical support.

5.2 Evaluation Task

To evaluate our scaffolding algorithms, we began by splitting the combined dataset into two blocks: (1) a scaffold corpus, and (2) an evaluation corpus. The agent's task was to predict the correct follow-on sentence for each dialog in the evaluation corpus.

Table 1: Dataset Statistics.

Chit-Chat	
Number of Words	$8,433,086$
Number of Turns	$483,112$
Average Length of Turns	88.86
Vocabulary Size	$85,952$
Daily Dialog	
Number of Words	$3,449,782$
Number of Turns	$243,520$
Average Length of Turns	62.85
Vocabulary Size	$26,116$
Reddit	
Number of Words	$33,847,503$
Number of Turns	$966,400$
Average Length of Turns	194.73
Vocabulary Size	$434,539$
Ubuntu	
Number of Words	$15,696,635$
Number of Turns	$966,400$
Average Length of Turns	88.33
Vocabulary Size	$145,594$

To set up this task, we first needed to standardize the formats of the datasets. Chit-Chat, Daily Dialog, and the Ubuntu Dialogue Corpus are all two-partner conversations. The Reddit data has a tree-like structure with an original post at the top, initial comments responding to the original post, more comments responding to the initial comments, and so forth. These threads of comments are not necessarily conversations between the same two users, as any user could post a comment in any thread. In standardizing the data, we chose to ignore distinctions between actual users and flatten the tree so that any Reddit thread was treated as a two-partner conversation between some speaker A and another speaker B.

Because some datasets, like Chit-Chat and the Ubuntu Dialogue Corpus, had many turns per conversation, we windowed our original data to create a sequence of shorter conversations with smaller dialog contexts. We chose a window size of four and a stride of one. Hence, if our original conversation had six turns, the windowed data would have three new

[3] https://github.com/BYU-PCCL/chitchat-dataset

[4] https://aclanthology.coli.uni-saarland.de/papers/I17-1099/i17-1099

[5] http://files.pushshift.io/reddit/

[6] https://www.kaggle.com/rtatman/ubuntu-dialogue-corpus

[7] Due to the massive size of Reddit, we only used a subset of the comments and posts from June 2014 to November 2014.

conversations with four turns each. We then set aside 3,311 conversations (about 5% of the smallest corpus) from each dataset to create the evaluation corpus, with the rest used as scaffolding.

Finally, the evaluation corpus was used to create a sequence of 13,244 windowed conversations. Each dialog from this evaluation set was paired with six candidate responses: (a) the correct follow-on sentence for the given dialog history, and (b) five distractors randomly chosen from the same text corpus as the correct answer. The scaffolding algorithms in Section 3.2, along with several baselines described in Section 5.3, were tasked with identifying the true response.

5.3 Baselines

We selected three baselines to compare against our candidate response scoring algorithms, our objective being to determine whether performance improves when the analogical structure of the embedding space is taken into consideration.

Naive Nearest
This algorithm is a non-analogical companion to the Naive Analogy algorithm depicted in Figure 2. Rather than calculating the ideal response as $d_1 = c + b_i - a_i$, the naive-nearest algorithm calculates $d_1 = b_i$. In other words, the Naive Nearest algorithm ignores the analogical nature of language by assuming that the successor to the best context match represents an optimal response, even if the contexts do not match exactly.

Approximate Nearest Neighbor (ANN)
This algorithm implements an Approximate Nearest Neighbor scoring strategy. Its ideal target point is calculated in the same way as the flow vectors algorithm, but with $d_1 = 1/n \sum b_i$. The analogical structure of the embedding space is ignored, and the algorithm instead orients itself based on the successor utterances extracted from the scaffold corpus.

Random
This baseline randomly selects one of the candidate responses without reference to the dialog history.

5.4 Results

Experimental results are shown in Table 2. Since no data was available on the relative ranking of the distractor sentences, we chose to evaluate our experimental results via response accuracy rather than via mean reciprocal rank or other metrics.

With a response accuracy of 68.07%, the scattershot algorithm shows a clear advantage over all other variants, outperforming the nearest baseline by 3.11%. We hypothesize that this is because scattershot takes the nondeterministic nature of language into account, allowing the scaffolding algorithm to select a candidate that most closely matches one of many possible valid responses.

Table 2: Algorithm accuracy on a response prioritization task with 13,244 distinct conversations. These experiments used a context length $n = 2$ and considered the $k = 5$ best context matches when calculating target point locations.

scoring method	accuracy
our algorithms	
flow vectors	62.47%
scattershot	**68.07%**
naive-analogy	62.29%
baselines	
naive-nearest	58.97%
ANN classifier	64.96%
random	16.06%

It is also useful to compare the naive-analogy algorithm with the naive-nearest baseline. These two algorithms are identical except for their analogical content. Our results show that leveraging the inherent analogical properties of the embedding space results in an overall accuracy improvement of 3.32%.

5.5 Analysis

A key insight to be taken from our results is the observation that the average of several valid responses cannot be assumed to *also* represent a valid response. This amounts to taking the centroid of a (possibly oddly-shaped) manifold, resulting in a vector that might easily represent a nonsense sentence.

For example, the question “Have you seen any good movies lately?” can be appropriately answered by statements including “Yes”, “No”, “I saw Iron Man last night”, and “Are you crazy? You know I’m afraid of theaters.” These disparate responses, all valid with respect to the dialog context, nevertheless vary widely in terms of sentence length, sentiment, vocabulary, and dialog mode. There is no reason to assume that a sentence which represents the average of all those factors will have any relevance to the dialog history, or that it will even be grammatically correct.

This observation explains the empirical success of the scattershot algorithm over its competitors. The naive-nearest and naive-analogy algorithms take only a single context match into consideration, and therefore function on limited information. The scattershot, flow vector and ANN algorithms all take many context matches into consideration, but while the flow vector and ANN algorithms operate by averaging multiple ‘ideal’ responses, the scattershot algorithm

```
> Hello.
Hey!
> How are you doing today?
How are you?
> I am good.  How has your day been so far?
I've been pretty good, I was sick earlier, but better now.
> Oh man, that stinks.  What were you sick with?  I am glad
you are feeling better though!
I had like ten servings of it.
> Uh oh.  That sounds like a lot.  Was it good food?
I had such tasty food too.
> Yum.  What is your favorite food?
yum sounds so good right now haha!  I love every fruit!  I
like rice and pickles too haha.  but like you I like just
about everything
> I love fruit too.  So, have you seen any good movies
lately?
Black panther is so good!!!
> Yeah I really liked it too.  I think that or, Thor
Ragnarok might be my favorite Marvel movie.
No , I haven't .  This is my first time to watch it .
> Oh yeah?
*It flips him off.*
> What are some of your favorite books?
It's a little bit of romance and comedy.
> Cool.  Well, I have to go.  It has been nice talking to
you!
Okay !  Bye !
```

Figure 5: Conversation generated using the scattershot scaffolding algorithm and Euclidean distance localization method, and distance threshold $\tau = 0.5$. Human-generated text is preceded by the $>$ symbol. In this example, the agent is able to maintain several fairly coherent dialog turns, then pivots appropriately to the topic of movies in response to user cues. It also successfully detects and responds to an indication that the conversation is over.

seeks a response that lies close to any one of the scaffold examples. This enables it to select the correct response candidate a higher percentage of the time.

6 RETRIEVAL FROM LARGE DATA REPOSITORIES

To explore the types of conversations produced by our system, we allowed a human to type the opening line of dialog with an agent executing the scattershot algorithm. The objective was to see whether a dialog agent choosing from among ca. 10,000 candidate responses could leverage the analogical structure of language in order to create a plausible conversation.

Typical conversations generated using this method are shown in Figures 5 and 6. Interestingly, although the 19,665 randomly-selected candidate utterances were drawn from all four conversational datasets, almost all of the ones chosen by the scattershot algorithm came from the Chit-Chat dataset. This suggests that the Chit-Chat dataset was an unusually good stylistic match for the informal conversation patterns used by the human chatter.

```
> Hey.  What's up?
Hey, whats up?
> Not much.  How is your day going so far?
Hi, it's good!
> Good to hear.  What are you working on?
I'm figuring out my budget .
> Oh that sounds fun.  What are you going to do after you
finish?
Well I plan to get into the accounting program
> Oh cool!  Do you like accounting?
I've thought about going into cardiology, that seems
interesting, but also I'm considering anesthesiology.
> Oh wow.  Those are both very different from accounting.
Are they hard?
semester for you and me.  Is your next semester seeming to
look like its going to be a little easier?
> I think this semester will be fun but a little
challenging.
That it does!  Even though you may not be getting
everything out of your classes now but oh well...;)
> Do you think you will have any time to go to see a movie
this semester?
That it does!  Even though you may not be getting
everything out of your classes now but oh well...;)
> So, where are you from originally?
Haha well I'm from Arkansas
> Oh cool.  I have never been there.  Do you like it?
Ya actually I did enjoy the shreveport one when I lived
there last year (2012 Mardi gras) but you're right, it's
almost nothing Compared to most of the NO ones.  Still very
enjoyable.  My of my friends lives right on the route so it
was nice having somewhere to go for it lol
> haha fun.  Do you like it here then?
Cool
> If you could vacation anywhere, where would it be?
well,its like 1 am here,i should get to bed, it was nice
talking to you, have a great day.
> Oh wow.  Goodnight.  Talk to you later.
Thanks.  You too!
```

Figure 6: Conversation generated using the scattershot scaffolding algorithm, the embedded concatenation localization method, and distance threshhold $\tau = 0.6$. Human-generated text is preceded by the $>$ symbol. In this example, even the τ threshhold is not sufficient to keep the agent from getting caught in a sentence repetition, however, it successfully switches to a new topic on the next utterance.

We permitted one augmentation to our algorithms for this experiment: Candidate responses that were too similar to the most recent statement in the dialog history were excluded from consideration.[8] This constitutes an extension at the sentence level of the traditional exclusion of source words when solving analogical queries via word embeddings (Mikolov et al., 2013b) (Gladkova et al., 2016). Without it, the scaffolding algorithm tends to select sentences that parrot or reflect the content of the dialog history rather than progressing to new topics.

7 CONCLUSIONS AND FUTURE WORK

As automated personal assistants become more prevalent, developers will need to strike a balance between control and spontaneity. It is important for automated agents to behave in unexpected, even surprising ways; otherwise they could not generalize from past experience in order to respond to unique queries from their users. At the same time, we do not want assistants who insult their users, make broadly offensive statements, or give inaccurate information.

The methods outlined in this paper provide a possible middle ground, allowing a scaffold corpus to define an overall personality or conversational style for the agent without directly restricting its possible utterances. In this paper, we have presented a scaffolding algorithm that uses pre-trained sentence embeddings to (a) leverage the inherent analogical properties of the embedding space and (b) account for the frequently non-deterministic nature of language while (c) encouraging responses that closely align with the scaffold corpus. Our scattershot algorithm is able to predict the correct follow-on sentence for a given dialog history with nearly 70% accuracy, outperforming both ANN and naive nearest-neighbor baselines.

Going forward, we imagine a possible future agent which generates responses via a neural architecture, but which has been trained to adhere as closely as possible to a scaffold corpus in its utterance patterns. Future work in this area should explore the possibility of neural dialog models that utilize a scaffold corpus during loss calculations, as well as the development of decoders that can render the target point directly into text. A comprehensive study of distance metrics should also be undertaken, as it is not certain that the *de facto* standards of Euclidean and cosine distance are the best possible heuristics for semantic similarity; L1 distance or correlation coefficients might be more effective.

[8] Similarity was defined as Euclidean distance $< \tau$, where τ is a hand-selected threshhold value.

ACKNOWLEDGEMENTS

We thank David Wingate and his students in the BYU Perception, Control and Cognition laboratory for their role in creating and hosting the Chit-Chat dataset.

REFERENCES

Al-Zubaide, H. and Issa, A. A. (2011). Ontbot: Ontology based chatbot. In *International Symposium on Innovations in Information and Communications Technology*, pages 7–12.

Banchs, R. E. and Li, H. (2012). Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Bartl, A. and Spanakis, G. (2017). A retrieval-based dialogue system utilizing utterance and context embeddings. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 1120–1125.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and individual Differences*, 67:97–102.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *CoRR*, abs/1803.11175.

Charras, F., Duplessis, G. D., Letard, V., Ligozat, A.-L., and Rosset, S. (2016). Comparing system-response retrieval models for open-domain and casual conversational agent. In *Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT@ IVA2016)*.

Cho, D. and Acquisti, A. (2013). The more social cues, the less trolling? an empirical study of online commenting behavior.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Dou, Z., Wei, W., and Wan, X. (2018). Improving word embeddings for antonym detection using thesauri and sentiwordnet. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 67–79. Springer.

Drozd, A., Gladkova, A., and Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on*

Computational Linguistics: Technical Papers, pages 3519–3530.

Dubuisson Duplessis, G., Charras, F., Letard, V., Ligozat, A.-L., and Rosset, S. (2017). Utterance Retrieval based on Recurrent Surface Text Patterns. In *39th European Conference on Information Retrieval*, Aberdeen, United Kingdom.

Fu, P., Lin, Z., Yuan, F., Wang, W., and Meng, D. (2018). Learning sentiment-specific word embedding via global sentiment representation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Fulda, N., Etchart, T., Myers, W., Ricks, D., Brown, Z., Szendre, J., Murdoch, B., Carr, A., and Wingate, D. (2018). Byu-eve: Mixed initiative dialog via structured knowledge graph traversal and conversational scaffolding. In *Proceedings of the 2018 Amazon Alexa Prize*.

Fulda, N., Ricks, D., Murdoch, B., and Wingate, D. (2017a). What can you do with a rock? affordance extraction via word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1039–1045.

Fulda, N., Tibbetts, N., Brown, Z., and Wingate, D. (2017b). Harvesting common-sense navigational knowledge for robotics from uncurated text corpora. In *Proceedings of the First Conference on Robot Learning (CoRL) - forthcoming*.

Gandhe, S. and Traum, D. (2013). Surface text based dialogue models for virtual humans. In *Proceedings of the SIGDIAL 2013 Conference*, pages 251–260.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Kim, J.-K., de Marneffe, M.-C., and Fosler-Lussier, E. (2016). Adjusting word embeddings with semantic intensity orders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 62–69.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought vectors. *CoRR*, abs/1506.06726.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv e-prints*, page arXiv:1710.03957.

Logeswaran, L. and Lee, H. (2018). An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *arXiv e-prints*, page arXiv:1506.08909.

Lowe, R. T., Pow, N., Serban, I. V., Charlin, L., Liu, C., and Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *D&D*, 8(1):31–65.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., tau Yih, W., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. Association for Computational Linguistics.

Nio, L., Sakti, S., Neubig, G., Yoshino, K., and Nakamura, S. (2016). Neural network approaches to dialog response retrieval and generation. *IEICE Transactions*, 99-D(10):2508–2517.

Patro, B. N., Kurmi, V. K., Kumar, S., and Namboodiri, V. P. (2018). Learning semantic sentence embeddings using sequential pair-wise discriminator. *arXiv preprint arXiv:1806.00807*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners.

Rainie, H. and Anderson, J. Q. (2017). *The future of free speech, trolls, anonymity and fake news online*.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Schneider, S. J., Kerwin, J., Frechtling, J., and Vivari, B. A. (2002). Characteristics of the discussion in online and face-to-face focus groups. *Social science computer review*, 20(1):31–42.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Thalenberg, B. (2016). Distinguishing antonyms from synonyms in vector space models of semantics. Technical report.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wu, Y., Wu, W., Xing, C., Zhou, M., and Li, Z. (2017). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. pages 496–505.

Zhu, X., Li, T., and De Melo, G. (2018). Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637.