A Neural Network Approach for Knowledge-Driven Response Generation

Pavlos Vougiouklis **Jonathon Hare** **Elena Simperl**
Electronics and Computer Science
University of Southampton
Southampton, UK
{pv1e13, jsh2, e.simperl}@ecs.soton.ac.uk

Abstract

We present a novel response generation system. The system assumes the hypothesis that participants in a conversation base their response not only on previous dialog utterances but also on their background knowledge. Our model is based on a Recurrent Neural Network (RNN) that is trained over concatenated sequences of comments, a Convolution Neural Network that is trained over Wikipedia sentences and a formulation that couples the two trained embeddings in a multimodal space. We create a dataset of aligned Wikipedia sentences and sequences of Reddit utterances, which we we use to train our model. Given a sequence of past utterances and a set of sentences that represent the background knowledge, our end-to-end learnable model is able to generate context-sensitive and knowledge-driven responses by leveraging the alignment of two different data sources. Our approach achieves up to 55% improvement in perplexity compared to purely sequential models based on RNNs that are trained only on sequences of utterances.

1 Introduction

Over the recent years, the level of users' engagement and participation in public conversations on social media, such as Twitter, Facebook and Reddit has substantially increased. As a result, we now have large amounts of conversation data that can be used to train computer programs to be proficient conversation participants. Automatic response generation could be immediately deployable in social media as an auto-complete response suggestion feature or a conversation stimulant that adjusts the participation interest in a dialogue thread (Ritter et al., 2011). It should also be beneficial in the development of Question-Answering systems, by enhancing their ability to generate human-like responses (Grishman, 1979).

Recent work on neural networks approaches shows their great potential at tackling a wide variety of Natural Language Processing (NLP) tasks (Bengio et al., 2003; Mikolov et al., 2010). Since a conversation can be perceived as a sequence of utterances, recent systems that are employed in the automatic response generation domain are based on Recurrent Neural Networks (RNNs) (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015), which are powerful sequence models. These systems base their generated response explicitly on a sequence of the most recent utterances of a conversation thread. Consequently, the sequence of characters, words, or comments, in a conversation, depending on the level of the model, is the only means with which these models achieve contextual-awareness, and in open-domain, realistic, situations it often proves inadequate (Vinyals and Le, 2015).

In this paper we address the challenge of context-sensitive response generation. We build a dataset that aligns knowledge from Wikipedia in the form of sentences with sequences of Reddit utterances. The dataset consists of sequences of comments and a number of Wikipedia sentences that were allocated randomly from the Wikipedia pages to which each sequence is aligned. The resultant dataset consists of $\sim$ 15k sequences of comments that are aligned with $\sim$ 75k Wikipedia sentences. We make the aligned corpus available at `github.com/pvougiou/Aligning-Reddit-and-Wikipedia`.

We propose a novel model that leverages this alignment of two different data sources. Our architecture is based on coupling an RNN using either Long Short-Term Memory (LSTM) cells (Hochreiter and

Schmidhuber, 1996) or Gated Recurrent Units (GRUs) (Cho et al., 2014) that processes each sequence of utterances word-by-word, and a Convolutional Neural Network (CNN) that extracts features from each set of sentences that corresponds to this sequence of utterances. We pre-train the CNN component (Kim, 2014) on a subset of the retrieved Wikipedia sentences in order to learn filters that are able to classify a sentence based on its referred topic. Our model assumes the hypothesis that each participant in a conversation bases their response not only on previous dialog utterances but also on their individual background knowledge. We use Wikipedia[1] as the source of our model's knowledge background and align Wikipedia pages and sequences of comments from Reddit[2] based on a predefined topic of discussion.

Our work is inspired by recent developments in the generation of textual summaries from visual data (Socher et al., 2014; Karpathy and Li, 2014). Our core insight stems from the idea that a system that is able to learn how to couple information from aligned datasets in order to produce a meaningful response, would be able to capture the context of a given conversation more accurately.

Our model achieves up to 55% improved perplexity compared to purely sequential equivalents. It should also be noted that our approach is domain independent; thus, it could be transferred out-of-box to a wide variety of conversation topics.

The structure of the paper is as follows. Section 2 discusses premises of our work regarding both automatic response generation and neural networks approaches for Natural Language Processing (NLP). Section 3 presents the components of the network. Section 4 describes the structure of the dataset. Section 5 discusses the experiments and the evaluation of the models. Section 6 summarises the contributions of the current work and outlines future plans.

2 Related Work

Since Bengio's introduction of neural networks in statistical language modelling (Bengio et al., 2003) and Mikolov's demonstration of the extreme effectiveness of RNNs for sequence modelling (Mikolov et al., 2010), neural-network-based implementations have been employed for a wide variety of NLP tasks. In order to sidestep the *exploding* and *vanishing gradients* training problem of RNNs (Bengio et al., 1994; Pascanu et al., 2012), multi-gated RNN variants, such as the GRU (Cho et al., 2014) and the LSTM (Hochreiter and Schmidhuber, 1996), have been proposed. Both GRUs and LSTMs have demonstrated state-of-the-art performance for many generative tasks, such as SMT (Cho et al., 2014; Sutskever et al., 2014), text (Graves, 2013) and image generation (Gregor et al., 2015).

Despite the fact that CNNs had been originally employed in the computer vision domain (LeCun et al., 1998), models based on the combination of the convolution operation with the classical Time-Delay Neural Network (TDNN) (Waibel et al., 1989) have proved effective on many NLP tasks, such as semantic parsing (Yih et al., 2014), Part-Of-Speech Tagging (POS) and Chunking (Collobert and Weston, 2008). Furthermore, sentence-level CNNs have been used in sentiment analysis and question type identification (Kalchbrenner et al., 2014; Kim, 2014).

The concept of a system capable of participating in human-computer conversations was initially proposed by Weizenbaum (Weizenbaum, 1966). Weizenbaum implemented ELIZA, a keyword-based program that set the basis for all the descendant *chatterbots*. In the years that followed, many template-based approaches (Isbell et al., 2000; Walker et al., 2003; Shaikh et al., 2010) have been suggested in the scientific literature, as a way of transforming the computer into a proficient conversation participant. However, these approaches usually adopt variants of the nearest-neighbour method to facilitate their response generation process from a number of limited sentence paradigms and, as a result, they are limited to specific topics or scenarios of conversation. Recently, models for Statistical Machine Translation have been used to generate short-length responses to a conversational incentive from Twitter utterances (Ritter et al., 2011). In the recent literature, RNNs have been used as the fundamental component of conversational response systems (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015). Even though these systems exhibited significant improvements over SMT-based methods (Ritter et al., 2011), they either adopt the length-restricted-messages paradigm or are trained on idealised dataset that undermines the

[1] http://www.wikipedia.com
[2] https://www.reddit.com

generation of responses in open domain realistic scenarios.

We propose a novel architecture for context-sensitive response generation. Our model is trained on a dataset that consists of realistic sequences of Reddit comments that aligned with sets of Wikipedia sentences. We use an RNN and a CNN components to process the sequence of comments and their corresponding set of sentences respectively and we introduce a learnable coupling formulation. The coupling formulation is inspired by the Multimodal RNN that generates textual description from visual data (Karpathy and Li, 2014). However, unlike Karpathy's approach, we do not allow the feature that is generated by the CNN component to diminish between distant timesteps (i.e. Section 3.3).

3 Our Model

Our task is to generate a context-sensitive response to a sequence of comments by incorporating background knowledge. The proposed model is based on the assumption that each participant in a conversation phrases their responses by taking into consideration both the past dialog utterances and their individual knowledge background. We train the model on a set of M sequences of Reddit comments and N summaries of Wikipedia pages that are related to the main discussed topic of a conversation. During training our models takes as an input a sequence of one-hot[3] vector representations of words $x_1, x_2, ..., x_T$ from a sequence of comments and a group of sentences S that is aligned with this sequence of utterances. We use a sentence-level CNN, which processes the group of sentences, in parallel with a word-level RNN that processes the sequence of comments in batches and propose a formulation that learns to couple the two networks to produce a meaningful response to the preceding comments. We experiment with two different commonly used RNN variants that are based on: (i) the LSTM cell and (ii) the GRU. We pre-train our CNN sentence model on a subset of the Wikipedia-sentences dataset in order for it to learn to classify a sentence based on the topic-keyword that was matched for its corresponding page acquisition.

Please note that since *bias* terms can be included in each weight-matrix multiplication (Bishop, 1995), they are not explicitly displayed in the equations that describe the models of this section.

3.1 Sentence Modelling

Models based on CNNs achieve their basic functionality by convolving a sequence of inputs with a set of filters in order to extract local features. We adopt the Convolutional Sentence Model from (Kim, 2014) and we expand it in order to meet our specific needs for a multi-class, rather than binary classification. Let $t_{1:l}$ the concatenation of the vectors of all the words that exist in a sentence s. A *narrow* type convolution operation with a filter $m \in \mathbb{R}^{k \times m}$ is applied to each m-gram in the sentence s in order to produce a *feature map* $\mathbf{c_{mf}} \in \mathbb{R}^{l-m+1}$ of features:

$$c_{mf_j} = tanh(m^T t_{j-m+1:j}) \ , \tag{1}$$

$$\mathbf{c_{mf}} = \begin{bmatrix} c_{mf_1} \\ \vdots \\ c_{mf_3} \end{bmatrix} \ , \tag{2}$$

with $l \geq m$. Shorter sentences are *padded* with zero vectors when necessary. The most relevant feature from each feature map is captured by applying the max-over-time pooling operation (Collobert and Weston, 2008). The consequent matrix is the result of concatenating the max values from each feature map that has been produced by applying an f number of m length filters over the sentence s. The network results in a fully-connected layer and a $softmax$ that carries out the classification of the sentences. The architecture of the sentence model is illustrated on the left side of Figure 1.

3.2 Sequence Modelling

We describe two commonly used RNN variants that are based on: (i) the LSTM cell and (ii) the GRU. We experiment with both of them in order to explore which one serves better the sequential-modelling

[3]Each x_t refers to the one-hot representation vector of a vocabulary token. One-hot is a vector that contains a 1 at the index of this particular x_t token in the vocabulary with all the other values set to zero.

needs of our full architecture.

Let $h_t^l \in \mathbb{R}^n$ be the aggregated output of a hidden unit at timestep $t = 1...T$ and layer depth $l = 1...L$. The vectors at zero layer depth, $h_t^0 = \mathbf{W}_{\mathbf{x}\rightarrow\mathbf{h}} x_t$, represent vectors that are given to the network as an input. The parameter matrix $\mathbf{W}_{\mathbf{x}\rightarrow\mathbf{h}}$ has dimensions $[|X|, n]$, where $|X|$ is the cardinality of all the potential one-hot input vectors. All the matrices that follow have dimension $[n, n]$.

3.2.1 Long Short-Term Memory

Our LSTM cells' architecture is adopted from (Zaremba and Sutskever, 2014):

$$in_t^l = sigm(\mathbf{W}_{\mathbf{in}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{in}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{3}$$

$$f_t^l = sigm(\mathbf{W}_{\mathbf{f}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{f}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{4}$$

$$cell_t^l = f_t^l \odot cell_{t-1}^l + in_t^l \odot tanh(\mathbf{W}_{\mathbf{cell}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{cell}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{5}$$

$$out_t^l = sigm(\mathbf{W}_{\mathbf{out}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{out}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{6}$$

$$h_t^l = out_t^l \odot tanh(cell_t^l) \ , \tag{7}$$

where in_t^l, f_t^l, out_t^l and $cell_t^l$ are the vectors at timestep t and layer depth l that correspond to the *input gate*, the *forget gate*, the *output gate* and the *cell* respectively.

3.2.2 Gated Recurrent Unit

The Gated Recurrent Unit was proposed as a less-complex implementation of the LSTM (Cho et al., 2014).

$$reset_t^l = sigm(\mathbf{W}_{\mathbf{reset}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{reset}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{8}$$

$$update_t^l = sigm(\mathbf{W}_{\mathbf{update}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{update}}^{\mathbf{l}} h_{t-1}^l) \ , \tag{9}$$

$$\widetilde{h}_t^l = tanh(\mathbf{W}_{\mathbf{in}}^{\mathbf{l}} h_t^{l-1} + \mathbf{W}_{\mathbf{h}\rightarrow\mathbf{h}}^{\mathbf{l}} (reset_t^l \odot h_{t-1}^l)) \ , \tag{10}$$

$$h_t^l = (1 - update_t^l) \odot h_{t-1}^l + update_t^l \odot \widetilde{h}_t^l \ , \tag{11}$$

$$\tag{12}$$

where $reset_t^l$, $update_t^l$ and $\widetilde{h}_t^l$ are the vectors at timestep t and layer depth l that represent the values of the *reset gate*, the *update gate* and the *hidden candidate* respectively.

3.3 Coupling

After the pre-training of the CNN is complete, and the fully-connected and $softmax$ layers are removed, the CNN is connected to the hidden units of the last layer L of the recurrent component. This is illustrated in Figure 1. The recurrent component is implemented with either LSTMs or GRUs. At each timestep, the RNN is processing a word from a sequence of comments and the CNN is extracting local features by convolving this sequence's corresponding sentences with groups of differently sized filters. The red-coloured edges in Figure 1 represent the learnable parameters during training.

The coupling formulation that follows is inspired by the Multimodal RNN that generates textual description from visual data (Karpathy et al., 2014). Since, we do not want to allow the effect of sentence features, which represent the background knowledge of our model, to diminish between distant timesteps we differentiate from Karpathy's approach; and instead of providing the feature that is generated by the CNN to the RNN only at the first timestep, we provide it at every timestep. Furthermore, Karpathy employs the *simple* RNN or Elman network (Elman, 1990) as the sequence modelling component of his architecture whereas we adopt multi-gated RNN variants.

It should be noted that in the equations that follow, the term $CNN_h \in \mathbb{R}^{\sum \mathbb{MF}}$ would refer to the output of the sentence-level CNN with its fully-connected and $softmax$ layers disconnected, where $\mathbb{MF}$ is the group of all the feature maps that are generated for each different filter size. During training the resultant

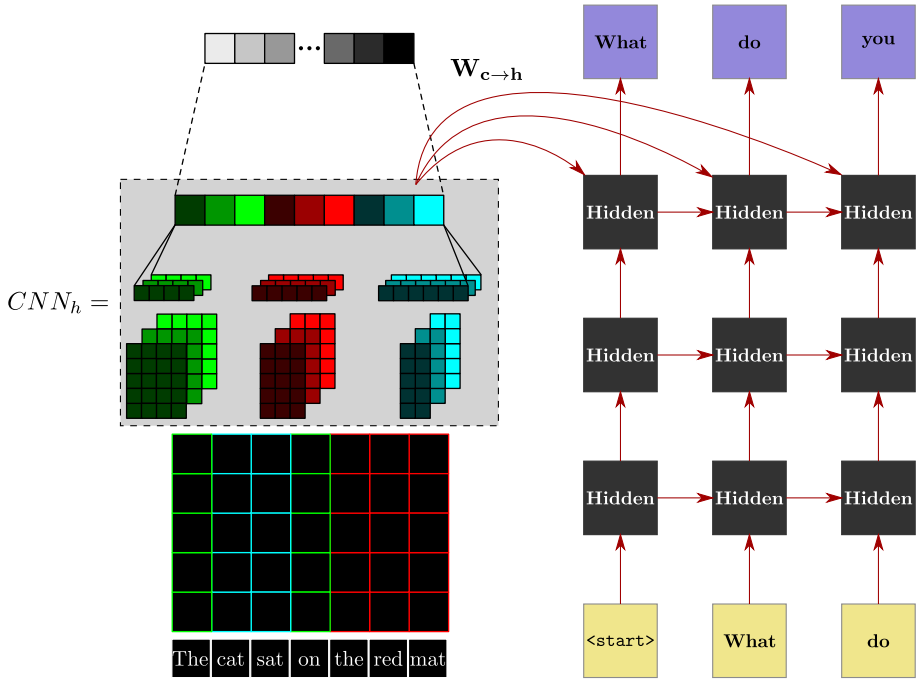

Figure 1: The architecture of our generative model. At each timestep, the RNN is processing a word from a sequence of comments and the CNN_h is extracting local features by convolving this sequence's corresponding sentences with a set of three differently sized filters. The red-coloured edges are the learnable parameters during training. Each comment in a sequence is augmented with start-of-comment `<start>` and end-of-comment `<end>` tokens.

embeddings, which are computed by the CNN_h processing a group of sentences S and the RNN variant processing the sequence of one-hot input word vectors $x_1, x_2, ..., x_T$ from the corresponding sequence of comments, are coupled in a hidden state $h_1, h_2, ..., h_T$. The prediction for the next word is computed by projecting this hidden state h_t to sequence of outputs $y_1, y_2, ..., y_T$:

$$c_S = \mathbf{W_{c \to h}} CNN_h(S) \ , \tag{13}$$

$$h_t = h_t^L \odot c_S \ , \tag{14}$$

$$y_t = softmax(\mathbf{W_y} h_t). \tag{15}$$

4 Dataset

We create a dataset[4] of aligned sentences from Wikipedia and sequences of utterances from Reddit. A shared, fixed, vocabulary was used for both data sources. We treat Wikipedia as a "cleaner" data source and we formed our vocabulary in the following manner. First, we included all the words that occur 2 or more times in the Wikipedia sentences. Subsequently, from the Reddit sequences of comments, we included any words that occur at least 3 times across both data sources. The resultant shared dictionary includes 56280 of the most frequent words. Every out-of-vocabulary word is represented by a special `NaN` token.

In constructing our dataset, our goal is to align Wikipedia sentences with sequences of comments from Reddit. We found that topics related to philosophy and literature are discussed on Reddit with the exchange of longer and more elaborative messages than the responses of the majority of conversational subjects on social media. A dataset that consists of long and detailed responses would provide more room for conversation incentives and would allow us to investigate the performance of our architectures against dialog exchanges with longer comments. We compiled a list of 35 predetermined topic-keywords

[4] `github.com/pvougiou/Aligning-Reddit-and-Wikipedia`

from the philosophical and literary domain. By utilising the `search` feature of both the Reddit API[5] and the MediaWiki API[6], we extracted: (i) sequences of comments from conversational threads most related to each keyword and (ii) the 300 Wikipedia pages most related to each keyword after carrying out Wikipedia's automatic disambiguation procedure. In order to increase the homogeneity of the dataset in terms of the length of both the sequences of comments and the sentences, we excluded sequences and sentences whose length exceeded: (i) $\overline{len} - \sigma = 1140$ and (ii) $\overline{len} - 2 \cdot \sigma = 54$ words respectively. The resultant dataset consists of 15460 sequences of Reddit comments and 75100 Wikipedia sentences.

4.1 Reddit

Reddit is absolved from the length-restricted-messages paradigm, facilitating the generation of longer and more meaningful responses. Furthermore, Reddit serves as an openly-available question-answering platform. Our research hypothesis is that a neural network trained on sequences of questions and their corresponding answers will be able to generate responses that escape the concept of daily-routine expressions, such as "good luck" or "have fun", and facilitate a playground for more detailed and descriptive dialogue utterances.

Each different conversation on Reddit starts with a user submitting a *parent-comment* on a subreddit. A sequence of utterances then succeeds this parent comment. Since we wanted to investigate how our model performs against long-term dependant dialog components, we set the depth of conversation to 5. Starting from the parent-comment, we follow the direction of the un-ordered tree of utterances until the fourth-level child-comment. If a comment (node) leads to n responses (children), we copy the observed sequence n times and for each sequence, we continue until the fourth response (leaf). Based on the above structural paradigm, we extracted sequences with at least four children-comments, of which we retained the only first four utterances along with the original parent-comment of the sequence. Note that each comment in a sequence is augmented with the respective start-of-comment `<start>` and end-of-comment `<end>` tokens.

Topic	**Reddit Sequence of Comments**	**Wikipedia Sentences**

Noam Chomsky	`<start>` Noam Chomsky: Bernie Sanders is Not a Radical. He has Mass Support for Positions on Healthcare & Taxes `<end>`
	`<start>` Funny, because Bernie Sanders's idol Eugene Debs ran against FDR `<end>`
	`<start>` Maybe Clinton will be FDR `<end>`
	`<start>` Watch out, Japanese. `<end>`
	`<start>` Japanese You misspelled Syrians `<end>`
		For Chomsky, there are minimalist questions but the answers can be framed in any theory.

		⋮
		Minimalism in structured writing or topic-based authoring is based on the ideas of John Millar Carroll.
		Minimalism is about reducing the interference of the information with the users sense-making process.
		An error, in fact, is the teachable moment that the content can exploit.

Table 1: Example of the alignment of our dataset. One sequence of comments is coupled with a set of sentences. The sentences are randomly allocated from the Wikipedia pages which have been extracted based on the same search term (**Noam Chomsky**) as the corresponding sequence.

4.2 Wikipedia

Wikipedia sentences are used as the knowledge background of our model. We chose to include only sentences from the Wikipedia summaries, since in preliminary experiments, we found that including all the textual material of a page introduces a lot of noise to our data. The 13410 Wikipedia summaries that matched the search criteria were split into sentences. Each sentence was labelled with the initial topic-keyword that was matched for its corresponding page acquisition. A subset of 30000 labelled sentences was used for pre-training the CNN component of our architecture.

[5] reddit.com/dev/api

[6] mediawiki.org/wiki/API:Main_page

4.3 Dataset Alignment

We choose to align each sequence of Reddit utterances with 20 Wikipedia sentences. Both the Wikipedia pages to which the sentences correspond and the sequence of comments have been extracted using the same search-term. An example of the structure of the dataset is displayed in Table 1.

5 Experiments

The full network was regularised by introducing a dropout (Zaremba et al., 2014) value of 0.4 to the non-recurrent connections between the last hidden state, h_t, and the softmax layer of the network. In order to avoid any potential exploding gradients training problems, we enforce an l_2 constraint on the gradients of the weights in order for them to be no greater than 5 (Sutskever et al., 2014).

- The CNN component is trained with narrow convolutional filters of widths 3, 4, 5 and 6, with 300 feature maps each. We use the rectifier as activation function. All of the parameters were initialised with a random uniform distribution between -0.1 and 0.1. The network was trained for 10 epochs using stochastic gradient descent with a learning rate of 0.2. We regularised the network by introducing a dropout (Hinton et al., 2012) value of 0.7 to the connections between the pooling and the softmax layer of the network.

- For the recurrent component of our networks, we use 2 layers of (i) 1000 LSTM cells and (ii) 1000 GRUs, resulting in approximately 16M and 12M recurrent connections respectively. All of the parameters are initialised with a random uniform distribution between -0.08 and 0.08. The networks were trained for 10 epochs, using stochastic gradient descent with a learning rate of 0.5. After the 7th epoch in the LSTM case and 3rd epoch in the GRU case, the learning rate was decayed by 0.2 every half epoch.

The dataset is split into training, validation and test with respective portions of 80, 10 and 10. A sample of responses that is generated by our proposed systems is shown in Table 2.

5.1 Experimental Results

Examples of responses that are generated by our proposed systems and their respective purely sequential equivalents are shown in Table 2. The sequences of comments and their corresponding sentences are sampled randomly from the test set. Our architectures learn to couple information that exists in the sequence of comments with knowledge that is contained in the Wikipedia sentences and is, potentially, related to context of those comments.

When a piece of information in the sequence of comments is successfully aligned with the content of its corresponding Wikipedia sentences a knowledgeable, context-sensitive response is generated. A representative example of this functionality is provided in the last sequence of comments in Table 2, where the context of the sequence is coupled with the fact that *Chomsky supported Bernie Sanders in the United States presidential election* (i.e. from the allocated to that sequence of Reddit utterances Wikipedia sentence: "In late 2015, Chomsky announced his support for Vermont U.S. senator Bernie Sanders in the upcoming 2016 United States presidential election."[7]). In case no information alignment is identified between the content of the sequence of comments and the Wikipedia sentences, the generation procedure is based almost explicitly on the sequence of utterances, and a response is generated in a similar to the purely sequential models' fashion.

5.2 Automatic Evaluation

We use perplexity on the test set to evaluate our proposed models against their purely sequential equivalents. Perplexity measures the cross-entropy between the predicted sequence of words and the actual, empirical, sequence of words. The results are illustrated in the top part of Table 3. Our proposed architectures achieve 55% and 45% improvement in perplexity compared to their respective purely LSTM- and GRU-based equivalents.

[7] en.wikipedia.org/wiki/Noam_Chomsky

Sequence of Comments	<start> My grandfather had a miniature set of all of Shakespeare's plays. It's 111 years old. <end> <start> Had? What happened to them? <end> <start> Less what happened to them, more what happened to him. <end> <start> I may have gone to your grandfather's estate sale. <end> <start> Literally as far away as you could be and stay in the same country. Also, we aren't having an estate sale. <end>	<start> Stephen Colbert & Stephen King <end> <start> This post made me realise that King played Bachman in Sons of Anarchy. He was fantastic in the role as well. <end> <start> Fun Fact: Stephen King writes under the pseudonym "Richard Bachman". <end> <start> Fun Fact: Richard NaN cause of death, per his creator, was "cancer of the pseudonym, a rare form of NaN." <end> <start> [Oh, here's a fun fact:] <end>
LSTM	<start> I don't. <end>	NaN.
GRU	<start> You are not so bad. <end>	<start> I meant that I was the one that was the first time to point. <end>
LSTM Coupled With ConvNet	<start> I was thinking the same. I saw it as a kid and I found Shakespeare and I was working in the head for a long time. <end>	<start> I was just thinking the same thing. <end>
GRU Coupled With ConvNet	<start> I don't know what I heard. <end>	<start> I remember searching for other people [spoiler] ("Apparently nothing is not "in" the books"). My personal favorite authors, not a whole story, but as a try to cause it to me. <end>
Sequence of Comments	<start> Can you write a six word story? According to legend, Ernest Hemingway once won a bet by writing a six word story comprised of these words: For sale. Baby shoes. Never worn. What's the best six word story you can write? Mine is, "She lied about having the procedure." <end> <start> I came; I saw; I conquered. <end> <start> "Veni, vidi, vici" said it better <end> <start> A better love story than twilight <end> <start> I'm fucking tired of this joke. <end>	<start> Noam Chomsky: The US would be recognized as a top terrorist state if international law was applied. <end> <start> You're not connecting the dots enough. The "regular" America gets cheap goods and oil through the bellicose actions of the "harsh" America. Now increasingly the actions of "harsh" America tend to benefit a tiny subset of the American people, but let's not pretend like our actions around the world don't have a multitude of practical effects on the American people as a whole. <end> <start> The thinking that if we didn't oppress people around the globe we wouldn't be able to maintain our lifestyle is misguided at best. <end> <start> So you're saying there's no economic benefit to our actions? I think you're mistaken. <end>
LSTM	<start> I think the "cause" is a good idea. <end>	<start> I don't think you are saying that it is a good idea. <end>
GRU	<start> No, it's not. <end>	<start> I would say that the "best" in the way we could have to do, and it is a big one. <end>
LSTM Coupled With ConvNet	<start> I don't think it's a mischaracterization to attribute this man's suicide to paranoia brought on by shady government surveillance. He was depressed, probably because of the holy few stories, the first step is to not like the world of his relationships, because he talks to the first one. <end>	<start> I don't think the majority of society has nothing to do with Sanders terrorism. <end>
GRU Coupled With ConvNet	<start> I think some of his memories was a family. <end>	<start> This is very radical I think Chomsky is an anarcho-syndicalist. An elementary linguistics student, he has no choice who is where he is. He does not need to be a "free market". He has to believe that the potential for American interests would be impossible, but they would be appropriate to charge what the outcome of American foreign interests is. <end>

Table 2: Sample of responses that are generated by our proposed systems and their sequential equivalents. The sequences of comments and their corresponding sentences are sampled randomly from the test set.

Model	LSTM	LSTM Coupled With ConvNet	GRU	GRU Coupled With ConvNet
Perplexity	4.301	**1.905**	3.749	2.051
Average Rating (σ)	**2.65** ($\pm$1.167)	2.4 ($\pm$1.27)	2.5 ($\pm$1.359)	**2.65** ($\pm$1.561)

Table 3: **Top:** Automatic evaluation with the perplexity metric on the test set. **Bottom:** Average rating of the responses that are generated by each model based on human evaluation.

5.3 Human Evaluation

Human evaluation was conducted using research students and staff from the School of Electronics and Computer Science of the University of Southampton. The evaluators were provided with a table of 10 randomly selected sequences of Reddit comments along with the response that is generated by our proposed models and their purely sequential equivalents. In order to simplify our task, we included only sequences of comments with a length less than 100 words. The name of the models to which each response corresponds were anonymised. The authors excluded themselves from this evaluation procedure. The evaluators were asked to rate each generated response from 1 to 5, with 1 indicating a very bad response, based on how well it fits the context of the corresponding sequence of comments.

Table 3 presents the average rating of each model's responses based on human evaluation. Even though our decision to apply a restriction over the length of the sequences of utterances, which were included in the human evaluation experiment, brings us in an agreement with literature that challenges the reliability of automatic evaluation methods, such as perplexity, in the domain of short-length responses (Ritter et al., 2011), we argue that an experiment at a larger scale, absolved from significant simplification choices, would demonstrate an alignment between the human judgements and the automatic evaluation results.

6 Conclusion

To the best of our knowledge this work constitutes the first attempt for building an end-to-end learnable system for automatic context-sensitive response generation by leveraging the alignment of two different data sources. We proposed a novel system that incorporates background knowledge in order to capture the context of a conversation and generate a meaningful response.

This paper made the following contributions: We built a dataset that aligns knowledge from Wikipedia in the form of sentences with sequences of Reddit utterances; and, we designed a neural network architecture that learns to couple information from different types of textual data in order to capture the context of a conversation and generate a meaningful response. Our approach achieved up to 55% improvement in perplexity compared to purely sequential models based on RNNs that are trained only on sequences of utterances. It should also be noted that despite the fact that our dataset focuses on the philosophical and literary domain, the design procedure could be transferred out-of-the-box to a great variety of domains.

Arguments could be made against the performance gain of our architectures against human evaluation. Based on the reported low performance of purely LSTM-based models on very long-term dependant datasets (Sutskever et al., 2014), we believe that an experiment at a larger scale without a restriction over the length of the sequences of utterances (Section 5.3) would emphasise the superiority of our approach.

We believe that further work on the coupling formulation that is proposed in Section 3.3 could provide additional improvements to the results of this work. An additional direction for future work could be the introduction of a complimentary, to the current procedure, task that would enhance the quality of the information alignment from the two data sources.

Acknowledgements

This research is partially supported by the Answering Questions using Web Data (WDAQua) project, a Marie Skłodowska-Curie Innovative Training Network, part of the Horizon 2020 programme. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

References

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.*, 5(2):157–166, March.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623.

Ralph Grishman. 1979. Response generation in question answering systems. In *Proceedings of the 17th Annual Meeting on Association for Computational Linguistics*, ACL '79, pages 99–101, Stroudsburg, PA, USA. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Sepp Hochreiter and Jrgen Schmidhuber. 1996. Bridging long time lags by weight guessing and "long short term memory". In *Spatiotemporal Models in Biological and Artificial Systems*, pages 65–72. IOS Press.

Charles Lee Isbell, Jr., Michael J. Kearns, Dave Kormann, Satinder P. Singh, and Peter Stone. 2000. Cobot in lambdamoo: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 36–41. AAAI Press.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Andrej Karpathy, Armand Joulin, and Fei Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1889–1897. Curran Associates, Inc.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, September 26-30, 2010*, pages 1045–1048.

Razvan Pascanu, Tomáš Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samira Shaikh, Tomek Strzalkowski, Sarah Taylor, and Nick Webb. 2010. Vca: An experiment with a multiparty virtual chat agent. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*, CDS '10, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China, July. Association for Computational Linguistics.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May–June. Association for Computational Linguistics.

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

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, Mar.

Marilyn Walker, Rashmi Prasad, and A Stent. 2003. A trainable generator for recommendations in multimodal dialog. In *Proceedings of EUROSPEECH*, pages 1697–1701.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January.

Wen-tau Yih, Xiaodong He, and Christopher Meek. 2014. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 643–648.

Wojciech Zaremba and Ilya Sutskever. 2014. Learning to execute. *CoRR*, abs/1410.4615.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.