NEURAL DISCOURSE MODELLING OF CONVERSATIONS

John M. Pierre, Mark Butler, Jacob Portnoff and Luis Aguilar

Voise Inc, San Francisco, USA

ABSTRACT

Deep neural networks have shown recent promise in many language-related tasks such as the modelling of conversations. We extend RNN-based sequence to sequence models to capture the long-range discourse across many turns of conversation. We perform a sensitivity analysis on how much additional context affects performance, and provide quantitative and qualitative evidence that these models can capture discourse relationships across multiple utterances. Our results show how adding an additional RNN layer for modelling discourse improves the quality of output utterances and providing more of the previous conversation as input also improves performance. By searching the generated outputs for specific discourse markers, we show how neural discourse models can exhibit increased coherence and cohesion in conversations.

KEYWORDS

Neural networks, Deep learning, Discourse, Conversation modelling

1. INTRODUCTION

Deep neural networks (DNNs) have been successful in modelling many aspects of natural language including word meanings [1][2], machine translation [3], syntactic parsing [4], language modelling [5] and image captioning [6]. Given sufficient training data, DNNs are highly accurate and can be trained end-to-end without the need for intermediate knowledge representations or explicit feature extraction. With recent interest in conversational user interfaces such as virtual assistants and chatbots, the application of DNNs to facilitate meaningful conversations is an area where more progress is needed. While sequence to sequence models based on recurrent neural networks (RNNs) have shown initial promise in creating intelligible conversations [7], it has been noted that more work is needed for these models to fully capture larger aspects of human communication including conversational goals, personas, consistency, context, and word knowledge.

Since discourse analysis considers language at the conversation-level, including its social and psychological context, it is a useful framework for guiding the extension of end-to-end neural conversational models. Drawing on concepts from discourse analysis such as *coherence* and *cohesion* [8], we can codify what makes conversations more intelligent in order to design more powerful neural models that reach beyond the sentence and utterance level. For example, by looking for features that indicate deixis, anaphora, and logical consequence in the machine-generated utterances we can benchmark the level of coherence and cohesion with the rest of the conversation, and then make improvements to models accordingly.

DOI: 10.5121/ijnlc.2016.5601

In the long run, if neural models can encode the long-range structure of conversations, they may be able to express conversational discourse similar to the way the human brain does, without the need for explicitly building formal representations of discourse theory into the model.

To that end, we explore RNN-based sequence to sequence architectures that can capture long-range relationships between multiple utterances in conversations and look at their ability to exhibit discourse relationships. Specifically, we look at 1) a baseline RNN encoder-decoder with attention mechanism and 2) a model with an additional discourse RNN that encodes a sequence of multiple utterances.

Our contributions are as follows:

- We examine two RNN models with attention mechanisms to model discourse relationships across different utterances that differ somewhat compared to what has been done before
- We carefully construct controlled experiments to study the relative merits of different models on multi-turn conversations
- We perform a sensitivity analysis on how the amount of context provided by previous utterances affects model performance
- We quantify how neural conversational models display coherence by measuring the prevalence of specific syntactical features indicative of deixis, anaphora, and logical consequence.

2. Related Work

Building on work done in machine translation, sequence to sequence models based on RNN encoder-decoders were initially applied to generate conversational outputs given a single previous message utterance as input [9][7]. In [10] several models were presented that included a “context” vector (for example representing another previous utterance) that was combined with the message utterance via various encoding strategies to initialize or bias a single decoder RNN. Some models have also included an additional RNN tier to capture the context of conversations. For example, [11] includes a hierarchical “context RNN” layer to summarize the state of a dialog, while [12] includes an RNN “intension network” to model conversation intension for dialogs involving two participants speaking in turn. Modelling the “persona” of the participants in a conversation by embedding each speaker into a *K*-dimensional embedding was shown to increase the consistency of conversations in [13].

Formal representations such as Rhetorical Structure Theory (RST) [14] have been developed to identify discourse structures in written text. Discourse parsing of cue phrases [15] and coherence modelling based on co-reference resolution of named-entities [16][17] have been applied to tasks such as summarization and text generation. Lexical chains [18] and narrative event chains [19] provide directed graph models of text coherence by looking at thesaurus relationships and subject-verb-temporal relationships, respectively. Recurrent convolutional neural networks have been used to classify utterances into discourse speech-act labels [20] and hierarchical LSTM models have been evaluated for generating coherent paragraphs in text documents [21].

Our aim is to develop end-to-end neural conversational models that exhibit awareness of discourse without needing a formal representation of discourse relationships.

3. MODELS

Since conversations are sequences of utterances and utterances are sequences of words, it is natural to use models based on an RNN encoder-decoder to predict the next utterance in the conversation given *N* previous utterances as source input. We compare two types of models: **seq2seq+A**, which applies an attention mechanism directly to the encoder hidden states, and **Nseq2seq+A**, which adds an additional RNN tier with its own attention mechanism to model discourse relationships between *N* input utterances.

In both cases the RNN decoder predicts the output utterance and the RNN encoder reads the sequence of words in each input utterance. The encoder and decoder each have their own vocabulary embeddings.

As in [4] we compute the attention vector at each decoder output time step *t* given an input sequence $(1,...,T_A)$ using:

$$\begin{aligned} u_i^t &= v^T tanh(W_1 h_i + W_2 d^t) \\ a_i^t &= softmax(u_i^t) \\ c^t &= \sum_{i=1}^{T_A} a_i^t h_i \end{aligned}$$

Where the vector v and matrices W_1, and W_2 are learned parameters. d^t is the decoder state at time t and is concatenated with c^t to make predictions and inform the next time step. In **seq2seq+A** the h_i are the hidden states of the encoder e_i, and for **Nseq2seq+A** they are the *N* hidden states of the discourse RNN (see Figure 1.) Therefore, in **seq2seq+A** the attention mechanism is applied at the word-level, while in **Nseq2seq+A** attention is applied at the utterance-level.

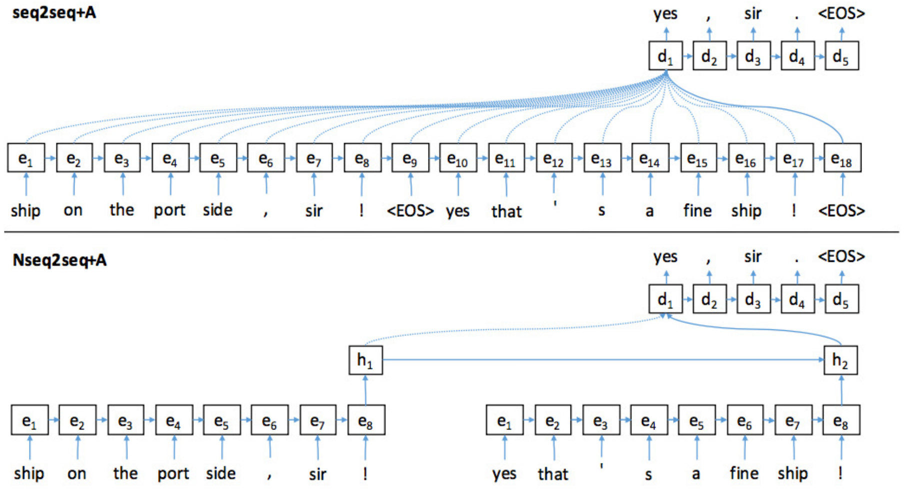
seq2seq+A
yes , sir . <EOS>
d1 d2 d3 d4 d5
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 e11 e12 e13 e14 e15 e16 e17 e18
ship on the port side , sir ! <EOS> yes that ' s a fine ship ! <EOS>
Nseq2seq+A
yes , sir . <EOS>
d1 d2 d3 d4 d5
h1
h2
e1 e2 e3 e4 e5 e6 e7 e8
ship on the port side , sir !
e1 e2 e3 e4 e5 e6 e7 e8
yes that ' s a fine ship !

Figure 1. Schematic of **seq2seq+A** and **Nseq2seq+A** models for multiple turns of conversation. An attention mechanism is applied either directly to the encoder RNN or to an intermediate discourse RNN.

3.1. Seq2seq+A

As a baseline starting point we use an attention mechanism to help model the discourse by a straightforward adaptation of the RNN encoder-decoder conversational model discussed in [7]. We join multiple source utterances using the *EOS* symbol as a delimiter, and feed them into the encoder RNN as a single input sequence. As in [3], we reversed the order of the tokens in each of the individual utterances but preserved the order of the conversation turns. The attention mechanism is able to make connections to any of the words used in earlier utterances as the decoder generates each word in the output response.

3.2. Nseq2seq+A

Since conversational threads are ordered sequences of utterances, it makes sense to extend an RNN encoder-decoder by adding another RNN tier to model the discourse as the turns of the conversation progress. Given *N* input utterances, the RNN encoder is applied to each utterance one at a time as shown in Figure 1 (with tokens fed in reverse order.) The output of the encoder from each of the input utterances forms *N* time step inputs for the discourse RNN. The attention mechanism is then applied to the *N* hidden states of the discourse RNN and fed into the decoder RNN. We also considered a model where the output of the encoder is also combined with the output of the discourse RNN and fed into the attention decoder, but found the purely hierarchical architecture performed better.

3.3. Learning

For each model, we chose identical optimizers, hyperparameters, etc. in our experiments in order to isolate the impact of specific differences in the network architecture, also taking computation times and available GPU resources into account. It would be straightforward to perform a grid search to tune hyperparameters, try LSTM cells, increase layers per RNN, etc. to further improve performance individually for each model beyond what we report here.

For each RNN we use one layer of Gated Recurrent Units (GRUs) with 512 hidden cells. Separate embeddings for the encoder and decoder, each with dimension 512 and vocabulary size of 40,000, are trained on-the-fly without using predefined word vectors.

We use a stochastic gradient descent (SGD) optimizer with *L2* norms clipped at *5.0*, an initial learning rate of *0.5*, and a learning rate decay factor of *0.99* is applied when needed. We trained with mini-batches of 64 randomly selected examples, and ran training for approximately 10 epochs until validation set loss converged.

4. EXPERIMENTS

We first present results comparing our neural discourse models trained on a large set of conversation threads based on the OpenSubtitles dataset [22]. We then examine how our models are able to produce outputs that indicate enhanced coherence by searching for discourse markers.

4.1. OpenSubtitles dataset

A large-scale dataset is important if we want to model all the variations and nuances of human language. From the OpenSubtitles corpus we created a training set and validation set with 3,642,856 and 911,128 conversation fragments, respectively (the training and validation sets consisted of *320M* and *80M* tokens, respectively). Each conversation fragment consists of 10 utterances from the previous lines of the movie dialog leading up to a target utterance. The main limitation of the OpenSubtitles dataset is that it is derived from closed caption style subtitles, which can be noisy, do not include labels for which actors are speaking in turn, and do not show conversation boundaries from different scenes.

We considered cleaner datasets such as the Ubuntu dialog corpus [23], Movie-DiC dialog corpus [24], and SubTle corpus [25] but found they all contained orders of magnitude fewer conversations and/or many fewer turns per conversation on average. Therefore, we found the size of the OpenSubtitles dataset outweighed the benefits of cleaner smaller datasets. This echoes a trend in neural networks where large noisy datasets tend to perform better than small clean datasets. The lack of a large-scale clean dataset of conversations is an open problem in the field.

4.2. Results

We compared models and performed a sensitivity analysis by varying the number of previous conversation turns fed into the encoder during training and evaluation.

In Table 1 we report the average perplexity (we use perplexity as our performance metric, because it is simple to compute and correlates with human judgements, though it has well-known limitations) on the validation set at convergence for each model. For *N=1,2,3* we found that **Nseq2seq+A** shows a modest but significant performance improvement over the baseline **seq2seq+A**. We only ran **Nseq2seq+A** on larger values of *N*, assuming it would continue to outperform.

Table 1. Results on OpenSubtitles dataset. Perplexity vs. number of previous conversation turns.

Previous conversation turns	**seq2seq+A**	**Nseq2seq+A**
N=1	13.84 ± 0.02	13.71 ± 0.03
N=2	13.49 ± 0.03	13.40 ± 0.04
N=3	13.44 ± 0.05	13.31 ± 0.03
N=5	-	13.14 ± 0.03
N=7	-	13.08 ± 0.03

In Figure 2 we show that increasing the amount of context from previous conversation turns significantly improves model performance, though there appear to be diminishing returns.

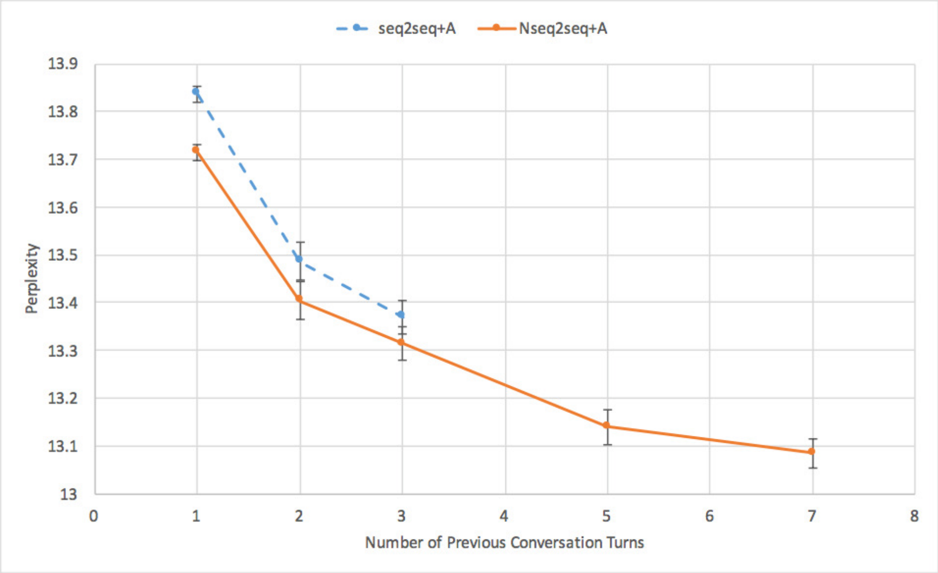

Figure 2. Sensitivity analysis of perplexity vs. number of previous conversations turns.

4.3. Discourse analysis

Since a large enough dataset tagged with crisp discourse relationships is not currently available, we seek a way to quantitatively compare relative levels of coherence and cohesion. As an alternative to a human-rated evaluation we performed simple text analysis to search for specific discourse markers [26] that indicate enhanced coherence in the decoder output as follows:

- **Deixis**: contains words or phrases (*here, there, then, now, later, this, that*) referring to previous context of place or time
- **Anaphora**: contains pronouns (*she, her, hers, he, him, his, they, them, their, theirs*) referring to entities mentioned in previous utterances
- **Logical consequence**: starts with a cue phrase (*so, after all, in addition, furthermore, therefore, thus, also, but, however, otherwise, although, if, then*) forming logical relations to previous utterances

In Table 2 we show how *N*, the number of previous conversation turns used as input, affects the likelihood that these discourse markers appear in the decoder output. The percentage of output utterances containing discourse markers related to deixis, anaphora, and logical consequence are reported from a sample of 100,000 validation set examples.

Table 2. Discourse analysis of **Nseq2seq+A** decoder output. Likelihood of discourse markers vs. number of previous conversation turns used as input.

Previous conversation turns	**N=1**	**N=2**	**N=3**	**N=5**	**N=7**
deixis	4.0%	3.4%	16.3%	5.1%	5.0%
anaphora	4.4%	6.1%	9.9%	7.2%	9.3%
logical consequence	0.03%	0.05%	0.08%	0.34%	0.12%

In general, we find that more context leads to a higher likelihood of discourse markers indicating that long-range discourse relationships are indeed being modelled. The results show a potentially interesting sensitivity to the value of *N*, require further study, and are likely to be dependent on different conversational styles and domains.

4.4. Examples

In Table 3 we show a few examples comparing decoder outputs of the **Nseq2seq+A** model using either 1 or 5 previous conversation turns as input. Qualitatively we can see that this neural discourse model is capable of producing increased cohesion when provided with more context.

Table 3. Example responses comparing **Nseq2seq+A** with either 1 or 5 previous conversation turns as input. Output words exhibiting cohesion with previous conversation turns are indicated in bold.

Previous Lines of Conversation Input	**Decoder Output**
can ' t see a number plate , even .	*i ' m sorry .*
then there ' s this . five streets away from UNK ' s flat , it ' s the night we broke in . it ' s him , it ' s the man we disturbed . we can ' t identify him from this . can ' t see a number plate , even .	*we ' re not going to get rid of **him** .*
but he was such a puppy !	*i was so scared .*
how time flies ! it only seems like yesterday since i put 00 candles on his birthday cake . he was such a tall boy ! nearly six feet . but he was such a puppy !	*he was a **young man** !*
now just two months ago right here	*and the other ?*
would you swear to that ? yes , sir , we swear to that . there you are , an UNK alibi . serena robbins , you come down here ! now just two months ago right here	***we ' re** on the **way down** .*

5. CONCLUSIONS

We studied neural discourse models that can capture long distance relationships between features found in different utterances of a conversation. We found that a model with an additional discourse RNN outperforms the baseline RNN encoder-decoder with an attention mechanism. Our results indicate that providing more context from previous utterances improves model performance up to a point. Qualitative examples illustrate how the discourse RNN produces increased coherence and cohesion with the rest of the conversation, while quantitative results based on text mining of discourse markers show that the amount of deixis, anaphora, and logical consequence found in the decoder output can be sensitive to the size of the context window.

In future work, it will be interesting to train discourse models on even larger corpora and compare conversations in different domains. By examining the attention weights, it should be possible to study what discourse markers the models are ``paying attention to'' and possibly provide a powerful new tool for analysing discourse relationships. By applying multi-task sequence to sequence learning techniques as in [27] we may be able to combine the conversational modelling task with other tasks such as discourse parsing and/or world knowledge modelling achieve better overall model performance. Not just for conversations, neural discourse modelling could also be applied to written text documents in domains with strong patterns of discourse such as news, legal, healthcare.

REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, (2013) “Efficient estimation of word representations in vector space”, arXiv preprint arXiv:1301.3781.

[2] Q.V. Le and T. Mikolov, (2014) “Distributed Representations of Sentences and Documents”, ICML.

[3] I. Sutskever, O. Vinyals, and Q.V. Le, (2014) “Sequence to sequence learning with neural networks”, Advances in neural information processing systems.

[4] O. Vinyals Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, (2015) “Grammar as a foreign language”, Advances in Neural Information Processing Systems.

[5] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, (2016) “Exploring the limits of language modeling”, arXiv preprint arXiv:1602.02410.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, (2015) “Show and tell: A neural image caption generator”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] O. Vinyals and Q.V. Le, (2015) “A neural conversational model”, arXiv preprint arXiv:1506.05869.

[8] M. Halliday and R. Hasan, (1976) “Cohesion in spoken and written English”, Longman Group Ltd.

[9] L. Shang, Z. Lu, H. Li, (2015) “Neural responding machine for short-text conversation”, arXiv preprint arXiv:1503.02364.

[10] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.Y. Nie, J. Gao, and B. Dolan, (2015) “A neural network approach to context-sensitive generation of conversational responses”, arXiv preprint arXiv:1506.06714.

[11] I.V. Serban, A Sordoni, Y Bengio, A Courville, and J. Pineau, (2016) “Hierarchical neural network generative models for movie dialogues”, arXiv preprint arXiv:1507.04808.

[12] K. Yao, G. Zweig, and B. Peng, (2015) “Attention with Intention for a Neural Network Conversation Model”, arXiv preprint arXiv:1510.08565.

[13] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, (2016) “A Persona-Based Neural Conversation Model”, arXiv preprint arXiv:1603.06155.

[14] W. Mann and S. Thompson, (1988) “Rhetorical structure theory: Toward a functional theory of text organization”, Text-Interdisciplinary Journal for the Study of Discourse 8.3.

[15] D. Marcu, (2000) “The theory and practice of discourse parsing and summarization”, MIT Press.

[16] R. Barzilay and M. Lapata, (2008) “Modeling local coherence: An entity-based approach”, Computational Linguistics 34.1.

[17] R. Kibble and R Power, (2004) “Optimizing referential coherence in text generation”, Computational Linguistics, MIT Press.

[18] J. Morris and G. Hirst, (1991) “Lexical cohesion computed by thesaural relations as an indicator of the structure of text”, Computational linguistics 17.1.

[19] N. Chambers and D. Jurafsky, (2008) “Unsupervised Learning of Narrative Event Chains”, ACL. Vol. 94305.

[20] N. Kalchbrenner and P. Blunsom, (2013) “Recurrent convolutional neural networks for discourse compositionality”, arXiv preprint arXiv:1306.3584.

[21] J. Li, MT Luong, and D. Jurafsky, (2015) “A hierarchical neural autoencoder for paragraphs and documents”, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.

[22] J. Tiedemann, (2009) “News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces”, N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia.

[23] R. Lowe, N. Pow, I. Serban, and J. Pineau, (2015) “The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems”, arXiv preprint arXiv:1506.08909.

[24] R.E. Banchs, (2012) “Movie-DiC: a movie dialogue corpus for research and development”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics.

[25] D. Ameixa, L. Coheur, and R.A. Redol, (2013) “From subtitles to human interactions: introducing the subtle corpus”, Tech. rep., INESC-ID.

[26] B. Fraser, (1999) “What are discourse markers? ”, Journal of Pragmatics 31.

[27] M.T. Luong, Q.V. Le, I. Sutskever, O. Vinyals, and Ł. Kaiser, (2014) “Multi-task sequence to sequence learning”, arXiv preprint arXiv:1511.06114.