DIALOG CONTEXT LANGUAGE MODELING WITH RECURRENT NEURAL NETWORKS

Bing Liu[1], Ian Lane[1,2]

[1]Electrical and Computer Engineering, Carnegie Mellon University
[2]Language Technologies Institute, Carnegie Mellon University

liubing@cmu.edu, lane@cmu.edu

ABSTRACT

In this work, we propose contextual language models that incorporate dialog level discourse information into language modeling. Previous works on contextual language model treat preceding utterances as a sequence of inputs, without considering dialog interactions. We design recurrent neural network (RNN) based contextual language models that specially track the interactions between speakers in a dialog. Experiment results on Switchboard Dialog Act Corpus show that the proposed model outperforms conventional single turn based RNN language model by 3.3% on perplexity. The proposed models also demonstrate advantageous performance over other competitive contextual language models.

Index Terms— RNNLM, contextual language model, dialog modeling, dialog act

1. INTRODUCTION

Language model plays an important role in many natural language processing systems, such as in automatic speech recognition [1, 2] and machine translation systems [3, 4]. Recurrent neural network (RNN) based models [5, 6] have recently shown success in language modeling, outperforming conventional n-gram based models. Long short-term memory [7, 8] is a widely used RNN variant for language modeling due to its superior performance in capturing longer term dependencies.

Conventional RNN based language model uses a hidden state to represent the summary of the preceding words in a sentence without considering context signals. Mikolov et al. proposed a context dependent RNN language model [9] by connecting a contextual vector to the RNN hidden state. This contextual vector is produced by applying Latent Dirichlet Allocation [10] on preceding text. Several other contextual language models were later proposed by using bag-of-word [11] and RNN methods [12] to learn larger context representation that beyond the target sentence.

The previously proposed contextual language models treat preceding sentences as a sequence of inputs, and they are suitable for document level context modeling. In dialog modeling, however, dialog interactions between speakers play an important role. Modeling utterances in a dialog as a sequence of inputs might not well capture the pauses, turn-taking, and grounding phenomena [13] in a dialog. In this work, we propose contextual RNN language models that specially track the interactions between speakers. We expect such models to generate better representations of the dialog context.

The remainder of the paper is organized as follows. In section 2, we introduce the background on contextual language modeling. In section 3, we describe the proposed dialog context language models. Section 4 discusses the evaluation procedures and results. Section 5 concludes the work.

2. BACKGROUND

2.1. RNN Language Model

A language model assigns a probability to a sequence of words $\mathbf{w} = (w_1, w_2, ..., w_T)$ following probability distribution. Using the chain rule, the likelihood of the word sequence $\mathbf{w}$ can be factorized as:

$$P(\mathbf{w}) = P(w_1, w_2, ..., w_T) = \prod_{t=1}^{T} P(w_t|w_{<t}) \quad (1)$$

At time step t, the system input is the embedding of the word at index t, and the system output is the probability distribution of the word at index $t+1$. The RNN hidden state h_t encodes the information of the word sequence up till current step:

$$h_t = \text{RNN}(h_{t-1}, w_t) \quad (2)$$

$$P(w_{t+1}|w_{<t+1}) = \text{softmax}(W_o h_t + b_o) \quad (3)$$

where W_o and b_o are the output layer weights and biases.

2.2. Contextual RNN Language Model

A number of methods have been proposed to introduce contextual information to the language model. Mikolov and Zweig [9] proposed a topic-conditioned RNNLM by introducing a contextual real-valued vector to RNN hidden state. The contextual vector was created by performing LDA [10] on preceding text. Wang and Cho [11] studied introducing

corpus-level discourse information into language modeling. A number of context representation methods were explored, including bag-of-words, sequence of bag-of-words, and sequence of bag-of-words with attention. Lin et al. [14] proposed using hierarchical RNN for document modeling. Comparing to using bag-of-words and sequence of bag-of-words for document context representation, using hierarchical RNN can better model the order of words in preceding text, at the cost of the increased computational complexity. These contextual language models focused on contextual information at the document level. Tran et al. [15] further proposed a contextual language model that consider information at inter-document level. They claimed that by utilizing the structural information from a tree-structured document set, language modeling performance was largely improved.

3. METHODS

The previously proposed contextual language models focus on applying context by encoding preceding text, without considering interactions in dialogs. These models may not be well suited for dialog language modeling, as they are not designed to capture dialog interactions, such as clarifications and confirmations. By making special design in learning dialog interactions, we expect the models to generate better representations of the dialog context, and thus lower perplexity of the target dialog turn or utterance.

In this section, we first explain the context dependent RNN language model that operates on utterance or turn level. Following that, we describe the two proposed contextual language models that utilize the dialog level context.

3.1. Context Dependent RNNLM

Let $\mathbf{D} = (\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_K)$ be a dialog that has K turns and involves two speakers. Each turn may have one or more utterances. The kth turn $\mathbf{U}_k = (w_1, w_2, ..., w_{T_k})$ is represented as a sequence of T_k words. Conditioning on information of the preceding text in the dialog, probability of the target turn $\mathbf{U}_k$ can be calculated as:

$$P(\mathbf{U}_k|\mathbf{U}_{<k}) = \prod_{t=1}^{T_k} P(w_t^{\mathbf{U}_k}|w_{<t}^{\mathbf{U}_k}, \mathbf{U}_{<k}) \quad (4)$$

where $\mathbf{U}_{<k}$ denotes all previous turns before $\mathbf{U}_k$, and $w_{<t}^{\mathbf{U}_k}$ denotes all previous words before the tth word in turn $\mathbf{U}_k$.

In context dependent RNN language model, the context vector c is connected to the RNN hidden state together with the input word embedding at each time step (Figure 1). This is similar to the context dependent RNN language model proposed in [9], other than that the context vector is not connected directly to the RNN output layer. With the additional context vector input c, the RNN state h_t is updated as:

$$h_t = \mathrm{RNN}(h_{t-1}, [w_t, c]) \quad (5)$$

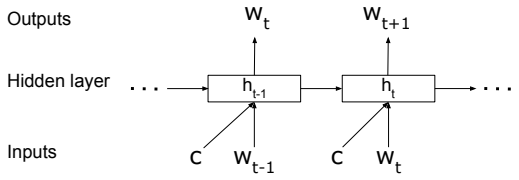

Fig. 1. Context dependent RNN language model.

3.2. Context Representations

In neural network based language models, the dialog context can be represented as a dense continuous vector. This context vector can be produced in a number of ways.

One simple approach is to use bag of word embeddings. However, bag of word embedding context representation does not take word order into consideration. An alternative approach is to use an RNN to read the preceding text. The last hidden state of the RNN encoder can be seen as the representation of the text and be used as the context vector for the next turn. To generate document level context representation, one may cascade all sentences in a document by removing the sentence boundaries. The last RNN hidden state of the previous utterance serves as the initial RNN state of the next utterance. As in [12], we refer to this model as DRNNLM. Alternatively, in the CCDCLM model proposed in [12], the last RNN hidden state of the previous utterance is fed to the RNN hidden state of the target utterance at each time step.

3.3. Interactive Dialog Context LM

The previously proposed contextual language models, such as DRNNLM and CCDCLM, treat dialog history as a sequence of inputs, without modeling dialog interactions. A dialog turn from one speaker may not only be a direct response to the other speaker's query, but also likely to be a continuation of his own previous statement. Thus, when modeling turn k in a dialog, we propose to connect the last RNN state of turn $k-2$ directly to the starting RNN state of turn k, instead of letting it to propagate through the RNN for turn $k-1$. The last RNN state of turn $k-1$ serves as the context vector to turn k, which is fed to turn k's RNN hidden state at each time step together with the word input. The model architecture is as shown in Figure 2. The context vector c and the initial RNN hidden state for the kth turn $h_0^{\mathbf{U}_k}$ are defined as:

$$c = h_{T_{k-1}}^{\mathbf{U}_{k-1}},\ h_0^{\mathbf{U}_k} = h_{T_{k-2}}^{\mathbf{U}_{k-2}} \quad (6)$$

where $h_{T_{k-1}}^{\mathbf{U}_{k-1}}$ represents the last RNN hidden state of turn $k-1$. This model also allows the context signal from previous turns to propagate through the network in fewer steps, which helps to reduce information loss along the propagation. We refer to this model as Interactive Dialog Context Language Model (IDCLM).

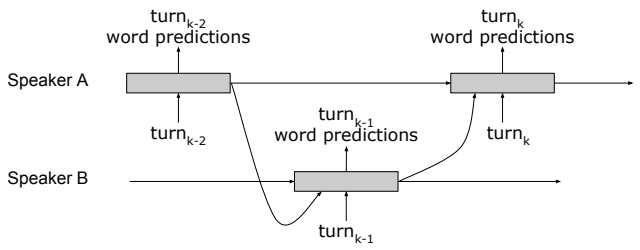

Fig. 2. Interactive Dialog Context Language Model (IDCLM).

3.4. External State Interactive Dialog Context LM

The propagation of dialog context can be seen as a series of updates of a hidden dialog context state along the growing dialog. IDCLM models this hidden dialog context state changes implicitly in the turn level RNN state. Such dialog context state updates can also be modeled in a separated RNN. As shown in the architecture in Figure 3, we use an external RNN to model the context changes explicitly. Input to the external state RNN is the vector representation of the previous dialog turns. The external state RNN output serves as the dialog context for next turn:

$$s_{k-1} = \mathrm{RNN}_{ES}(s_{k-2}, h_{T_{k-1}}^{\mathbf{U}_{k-1}}) \tag{7}$$

where s_{k-1} is the output of the external state RNN after the processing of turn $k-1$. The context vector c and the initial RNN hidden state for the kth turn $h_0^{\mathbf{U}_k}$ are then defined as:

$$c = s_{k-1},\ h_0^{\mathbf{U}_k} = h_{T_{k-2}}^{\mathbf{U}_{k-2}} \tag{8}$$

We refer to this model as External State Interactive Dialog Context Language Model (ESIDCLM).

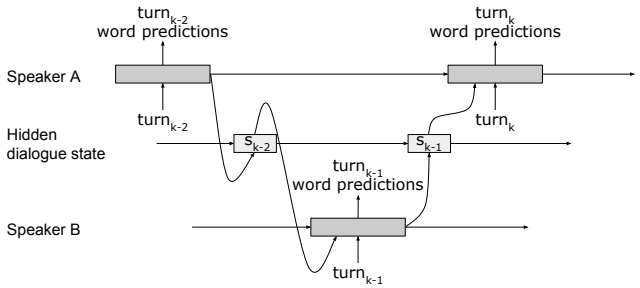

Fig. 3. External State Interactive Dialog Context Language Model (ESIDCLM).

Comparing to IDCLM, ESIDCLM releases the burden of turn level RNN by using an external RNN to model dialog context state changes. One drawback of ESIDCLM is that there are additional RNN model parameters to be learned during model training, which may make the model more prone to overfitting when training data size is limited.

4. EXPERIMENTS

4.1. Data Set

We use the Switchboard Dialog Act Corpus (SwDA)[1] in evaluating our contextual langauge models. The SwDA corpus extends the Switchboard-1 Telephone Speech Corpus with turn and utterance-level dialog act tags. The utterances are also tagged with part-of-speech (POS) tags. We split the data in folder sw00 to sw09 as training set, folder sw10 as test set, and folder sw11 to sw13 as validation set. The training, validation, and test sets contain 98.7K turns (190.0K utterances), 5.7K turns (11.3K utterances), and 11.9K turns (22.2K utterances) respectively. Maximum turn length is set to 160. The vocabulary is defined with the top frequent 10K words.

4.2. Baselines

We compare IDCLM and ESIDCLM to several baseline methods, including n-gram based model, single turn RNNLM, and various context dependent RNNLMs.

5-gram KN A 5-gram language model with modified Kneser-Ney smoothing [16].

Single-Turn-RNNLM Conventional RNNLM that operates on single turn level with no context information.

BoW-Context-RNNLM Contextual RNNLM with BoW representation of preceding text as context.

DRNNLM Contextual RNNLM with turn level context vector connected to initial RNN state of the target turn.

CCDCLM Contextual RNNLM with turn level context vector connected to RNN hidden state of the target turn at each time step. We implement this model following the design in [12].

In order to investigate the potential performance gain that can be achieved by introducing context, we also compare the proposed methods to RNNLMs that use true dialog act tags as context. Although human labeled dialog act might not be the best option for modeling the dialog context state, it provides a reasonable estimation of the best gain that can be achieved by introducing linguistic context. The dialog act sequence is modeled by a separated RNN, similar to the external state RNN used in ESIDCLM. We refer to this model as Dialog Act Context Language Model (DACLM).

DACLM RNNLM with true dialog act context vector connected to RNN state of the target turn at each time step.

4.3. Model Configuration and Training

In this work, we use LSTM cell [7] as the basic RNN unit for its stronger capability in capturing long-range dependencies in a word sequence comparing to simple RNN. We use pretrained word vectors [17] that are trained on Google News dataset to initialize the word embeddings. These word embeddings are fine-tuned during model training. We conduct

[1] http://compprag.christopherpotts.net/swda.html

mini-batch training using Adam optimization method following the suggested parameter setup in [18]. Maximum norm is set to 5 for gradient clipping . For regularization, we apply dropout ($p = 0.8$) on the non-recurrent connections [19] of LSTM. In addition, we apply L_2 regularization ($\lambda = 10^{-4}$) on the weights and biases of the RNN output layer.

4.4. Results and Analysis

The experiment results on language modeling perplexity for models using different dialog turn size are shown in Table 1. K value indicates the number of turns in the dialog. Perplexity is calculated on the last turn, with preceding turns used as context to the model.

Table 1. Language modeling perplexities on SwDA corpus with various dialog context turn sizes (K).

Model	K=1	K=2	K=3	K=5
5-gram KN	65.7	-	-	-
Single-Turn-RNNLM	60.4	-	-	-
BoW-Context-RNNLM	-	59.6	59.2	58.9
DRNNLM	-	60.1	58.6	59.1
CCDCLM	-	63.9	61.4	62.2
IDCLM	-	-	58.8	58.6
ESIDCLM	-	-	**58.4**	**58.5**
DACLM	-	58.2	57.9	58.0

As can be seen from the results, all RNN based models outperform the n-gram model by large margin. The BoW-Context-RNNLM and DRNNLM beat the Single-Turn-RNNLM consistently. Our implementation of the context dependent CCDCLM performs worse than Single-Turn-RNNLM. This might due to fact that the target turn word prediction depends too much on the previous turn context vector, which connects directly to the hidden state of current turn RNN at each time step. The model performance on training set might not generalize well during inference given the limited size of the training set.

The proposed IDCLM and ESIDCLM beat the single turn RNNLM consistently under different context turn sizes. ESIDCLM shows the best language modeling performance under dialog turn size of 3 and 5, outperforming IDCLM by a small margin. IDCLM beats all baseline models when using dialog turn size of 5, and produces slightly worse perplexity than DRNNLM when using dialog turn size of 3.

To analyze the best potential gain that may be achieved by introducing linguistic context, we compare the proposed contextual models to DACLM, the model that uses true dialog act history for dialog context modeling. As shown in Table 1, the gap between our proposed models and DACLM is not wide. This gives a positive hint that the proposed contextual models may implicitly capture the dialog context state changes.

For fine-grained analyses of the model performance, we further compute the test set perplexity per POS tag and per dialog act tag. We selected the most frequent POS tags and dialog act tags in SwDA corpus, and report the tag based perplexity relative changes (%) of the proposed models comparing to Single-Turn-RNNLM. A negative number indicates performance gain.

Table 2. Perplexity relative change (%) per POS tag

POS Tag	IDCLM	ESIDCLM	DACLM
PRP	-16.8	-5.8	-10.1
IN	-2.0	-5.5	-1.8
RB	-4.1	-8.9	-4.3
NN	13.4	8.1	2.3
UH	-0.4	7.7	-9.7

Table 2 shows the model perplexity per POS tag. All the three context dependent models produce consistent performance gain over the Single-Turn-RNNLM for pronouns, prepositions, and adverbs, with pronouns having the largest perplexity improvement. However, the proposed contextual models are less effective in capturing nouns. This suggests that the proposed contextual RNN language models exploit the context to achieve superior prediction on certain but not all POS types. Further exploration on the model design is required if we want to better capture words of a specific type.

Table 3. Perplexity relative change (%) per dialog act tag.

DA Tag	IDCLM	ESIDCLM	DACLM
Statement-non-opinion	-1.8	-0.5	-1.6
Acknowledge	-2.6	11.4	-16.3
Statement-opinion	4.9	-0.9	-1.0
Agree/Accept	14.7	2.7	-15.1
Appreciation	0.7	-3.8	-6.5

For the dialog act tag based results in Table 3, the three contextual models show consistent performance gain on Statement-non-opinion type utterances. The perplexity changes for other dialog act tags vary for different models.

5. CONCLUSIONS

In this work, we propose two dialog context language models that with special design to model dialog interactions. Our evaluation results on Switchboard Dialog Act Corpus show that the proposed model outperform conventional RNN language model by 3.3%. The proposed models also illustrate advantageous performance over several competitive contextual language models. Perplexity of the proposed dialog context language models is higher than that of the model using true dialog act tags as context by a small margin. This indicates that the proposed model may implicitly capture the dialog context state for language modeling.

6. REFERENCES

[1] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech recognition," 1993.

[2] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[3] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[5] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model.," in *Interspeech*, 2010, vol. 2, p. 3.

[6] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5528–5531.

[7] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "Lstm neural networks for language modeling.," in *Interspeech*, 2012, pp. 194–197.

[9] Tomas Mikolov and Geoffrey Zweig, "Context dependent recurrent neural network language model.," in *SLT*, 2012, pp. 234–239.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[11] Tian Wang and Kyunghyun Cho, "Larger-context language modelling," *arXiv preprint arXiv:1511.03729*, 2015.

[12] Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein, "Document context language models," *arXiv preprint arXiv:1511.03962*, 2015.

[13] Herbert H Clark and Susan E Brennan, "Grounding in communication," *Perspectives on socially shared cognition*, vol. 13, no. 1991, pp. 127–149, 1991.

[14] Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li, "Hierarchical recurrent neural network for document modeling," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 899–907.

[15] Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari, "Inter-document contextual language model," in *Proceedings of NAACL-HLT*, 2016, pp. 762–766.

[16] Stanley F Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.

[17] T Mikolov and J Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.