Session-level Language Modeling for Conversational Speech

Wayne Xiong **Lingfeng Wu** **Jun Zhang**
Microsoft, Bellevue, WA, USA
{weixi,lingfw,junzh}@microsoft.com

Andreas Stolcke
Microsoft, Sunnyvale, CA, USA
anstolck@microsoft.com

Abstract

We propose to generalize language models for conversational speech recognition to allow them to operate across utterance boundaries and speaker changes, thereby capturing conversation-level phenomena such as adjacency pairs, lexical entrainment, and topical coherence. The model consists of a long-short-term memory (LSTM) recurrent network that reads the entire word-level history of a conversation, as well as information about turn taking and speaker overlap, in order to predict each next word. The model is applied in a rescoring framework, where the word history prior to the current utterance is approximated with preliminary recognition results. In experiments in the conversational telephone speech domain (Switchboard) we find that such a model gives substantial perplexity reductions over a standard LSTM-LM with utterance scope, as well as improvements in word error rate.

1 Introduction

Over the past decade the state of the art in language modeling has shifted from N-gram models to feed-forward networks (Bengio et al., 2006), and then to recurrent neural networks (RNNs) that read a list of words sequentially and predict the next word at each position. Starting with standard recurrent networks (Mikolov et al., 2010) the sequential modeling approach was later improved using the long-short-term memory (LSTM) architecture of (Hochreiter and Schmidhuber, 1997) for further gains (Sundermeyer et al., 2012; Medennikov et al., 2016; Xiong et al., 2017). RNN models give two fundamental advantages over the old N-gram framework. First, the continuous-space embedding of word identities allows word similarities to be exploited for generalization (Bengio et al., 2006; Mikolov et al., 2013). Second, the recurrent architecture allows, in principle at least, an unlimited history to condition the prediction of next words.

The potential advantage of unlimited history, however, is not commonly used to its full benefit, since the language model (LM) is typically "reset" at the start of each utterance in current state-of-the-art recognition systems (Saon et al., 2017; Xiong et al., 2018). This presumes that each utterance is independent of the others, and clearly violates what we know about how language and conversation works, as discussed in the next section. Consequently, there have been many proposals to inject information from a longer context into standard LM architectures, going back to N-gram models (Bellegarda, 2004), or to generalize N-grams LMs to operate across utterance boundaries and speakers (Ji and Bilmes, 2004). Based on the RNN framework, (Mikolov and Zweig, 2012) proposed augmenting network inputs with a more slowly varying context vector that would encode longer-range properties of the history, such as a latent semantic indexing vector. The problem with these approaches is that the modeler has to make design decisions about how to encapsulate contextual information as network inputs. Therefore, our approach here is to simply provide the entire conversation history as input to a standard LSTM-LM, and let the network learn the information that is relevant to next-word prediction.

We start by discussing linguistic phenomena that could potentially help in conversational LM (Section 2), followed by a description of the LSTM model we propose to capture them (Section 3). Section 4 describes the data and recognition system we used to test our models, with results reported in Section 5. We end with conclusions and future directions.

2 Conversation-level Phenomena

Here we review a few of the conversation-level phenomena that could be used for predicting words from longer context. Perhaps the most widely studied effect is topical coherence, or the tendency of words that are semantically related to one or more underlying topics to appear together in the conversation. Consequently, topic-related words are bound to re-occur across utterances, or certain related words appear to trigger one another (such as "children" and "school"). This should be especially true for conversations in the Switchboard (and Fischer) corpora, which were collected by pairing up strangers to talk about a mutually agreeable topic.

Another phenomenon that could lead to words reoccurring is lexical entrainment (Brennan and Clark, 1996), or the tendency of conversants to adopt the same words and phrases. Entrainment can also apply to speaking style, so the use of common discourse particles, syntactic patterns (like question tags), or even disfluencies could be triggered across speakers.

Other phenomena operate more locally, but across speaker turn boundaries. Linguistic conversation analysis has long noted that utterance types come in *adjacency pairs* (Schegloff, 1968), with preferences for certain pairs over others (like a statement is preferentially followed by agreement rather than disagreement). Therefore, words in an utterance should be more predicable based on the previous utterance. In the past, this has been modeled by conditioning utterance words on an underlying dialog act label, which in turn is conditioned on adjacent dialog act labels via a dialog act grammar (Stolcke et al., 2000).

A good part of conversational behavior has to do with how turn-taking is negotiated (Sacks et al., 1974). Speakers use special discourse devices, such as backchannel words and pause fillers, to signal when they want to take the floor, or to signal that the other party should keep the floor. Conversants also anticipate the ends of turns and jump in before the other speaker is completely done, making for very efficient use of time. As a result of all of these mechanisms, a good portion of conversations consists of overlapping (simultaneous) speaking. It was shown (Shriberg et al., 2001) that such overlap locations can be partly predicted by word-based language models. This suggests reversing the modeling and using overlap (the timing of utterances) to help predict the words.

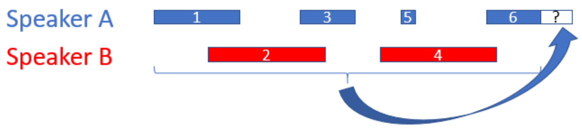

Figure 1: Use of conversation-level context in session-based LM. The utterance numbering shows how overlapping utterances are serialized (according to onset times).

3 Models

Our baseline language model is a standard LSTM that models utterances independently from one another, i.e., the history at the onset of each utterance is the start-of-sentence token. In fact, we used two version of this basic LSTM-LM:

- Word inputs encoded with one-hot vectors, combined with a jointly trained embedding layer
- Words encoded by multiple-hot vectors corresponding to the letter trigrams making up the words.

Both types of LSTM-LMs use three 1000-dimensional hidden layers with recurrence. The word embedding layer is also of size 1000, and the letter-trigram encoding has size 7190 (the number of unique trigrams in our vocabulary).

The main addition for session-level modeling is that the LSTM history consists of all the utterances preceding the current utterance, followed by all words in the current utterance preceding the word to be predicted. The preceding utterances are serialized in the order of their onset times, so that the flow of words within an utterance is not disrupted. The resulting total word history and next-word prediction is depicted in Figure 1. Information about utterance boundaries is encoded using a boundary tag, similar to the start-of-sentence token that is commonly used in LMs.

Several of the conversational phenomena described in Section 2 refer to turn-taking between speakers; to capture this in the model we augment the word input encoding with an extra bit that indicates whether a speaker change occurred. This bit is turned on only for the start-of-utterance token.

We also want to capture some information about utterance overlap, since, as described earlier, speech overlap interacts with word choice. Possible events to model would be overlap (exceedings a time threshold) at the starts and ends of utterances, or maybe a continuous measure of such overlaps. As a first proof of concept we chose to encode only one type of overlap, i.e., when the utterance in question is completely overlapped temporally by the other speaker's turn. This is typical of backchannel acknowledgments ("uh-huh") and short utterances that attempt to grab the floor ("um", "but"). Complete utterance overlap is also encoded by an additional input bit that is turned on for the start-of-utterance token.

4 Experiments

4.1 Recognition system

We used a single bidirectional LSTM acoustic model in experiments reported here, trained on the commonly used conversational telephone speech corpora (Switchboard, Fisher, CallHome English), estimating frame-level posterior probabilities for 9000 context-dependent phone units. The system decodes speech utterances using a 4-gram language model, generating lattices. These are then expanded to 500-best lists, which in turn are rescored using the various LMs.

The recognition system and the N-gram LM used in decoding have a vocabulary of 165k words, but the LSTM-LMs are trained on only the 38k words occurring at least twice in the in-domain conversational training data. Words outside of the LSTM-LM vocabulary are penalized in rescoring with a constant weight that is empirically optimized on the development set.

4.2 Data

Language model training uses the Switchboard-1, BBN Switchboard-2, Fisher, and English CallHome transcripts (about 23 million words in total) as well as the UW conversational Web corpus (Bulyko et al., 2003) for pre-training (see below). The N-gram LM used for N-best generation also includes the LDC Hub4 (Broadcast News) corpus. The Switchboard-1 and Switchboard-2 portions of the NIST 2002 CTS test set were used for tuning and development. Evaluation is carried out on the NIST 2000 CTS test set, consisting of Switchboard (SWB) and CallHome (CH) subsets.

As an expedient, we refrained from resegmenting utterances based on forced alignments of words, and instead use utterance boundaries as given in the available transcripts (corresponding to the audio segments used in acoustic training). Similarly, in testing, we use the presegmented utterances provided by NIST. No doubt there are inconsistencies in how the different corpora define utterance units, and a consistent, alignment-based resegmentation of all training and test data based on the durations nonspeech regions and/or lexical tagging might give improved results.

Table 1: Perplexities with session-based LSTM-LMs. The last two lines reflect use of errorful recognition output for preceding utterances.

Model inputs	devset SWB	test SWB	test CH
Utterance words, letter-3grams	48.90	44.56	54.57
+ session history words	38.86	36.81	44.31
+ speaker change	37.25	35.33	42.23
+ speaker overlap	37.09	35.12	42.02
Using recognized word histories			
single system	39.55	37.45	46.49
full system (Xiong et al., 2018)	39.41	37.29	45.99

4.3 Model training

All LSTM-LMs are trained using the Microsoft Cognitive Toolkit, or CNTK (Yu et al., 2014; Microsoft Research, 2016) on a Linux-based multi-GPU server farm. Training is parallelized using CNTK's distributed stochastic gradient descent (SGD) with 1-bit gradient quantization (Seide et al., 2014). We use the CNTK "FsAdaGrad" learning algorithm, which is an implementation of Adam (Kingma and Ba, 2015).

All LSTM-LMs are pretrained for one or two epochs on a large corpus of "conversational Web" data (Bulyko et al., 2003), followed by normal training to convergence on the in-domain data. Each utterance in the Web data is treated as a single session for purposes of session-based LM, i.e., the extra bits for speaker change and overlap are never turned on.

5 Results

When evaluating the session-based LMs on speech test data, the true utterance contexts are not known, and we must use hypothesized words for word histories preceding the current utterance. In our case, the histories were obtained using the output of our best recognition system, which uses a combination of acoustic models (Xiong et al., 2018), but excluding the session-based LM.[1] Per-

[1] We also omitted the final confusion network rescoring stage described in (Xiong et al., 2018).

Table 2: Recognition results with standard and session-based LSTM-LMs, measured by word error rates (WER).

Word encoding	Model	WER devset	WER test SWB	WER test CH
Letter 3gram	LSTM-LM	10.01	6.88	12.79
	Session LSTM-LM	9.67	6.81	12.54
	Session LSTM-LM, 2nd iteration	9.66	6.77	12.56
One-hot	LSTM-LM	9.81	6.89	13.02
	Session LSTM-LM	9.47	6.81	12.60
	Session LSTM-LM, 2nd iteration	9.50	6.83	12.73
Letter 3gram + One-hot	LSTM-LM	9.66	6.63	12.77
	Session LSTM-LM	9.28	6.52	12.34
	LSTM-LM + Session LSTM-LM	9.22	6.45	12.11

plexity was evaluated on reference transcripts, as is customary.

Table 1 shows the effect of session-level modeling and of optional model elements on perplexity, based on LSTMs using letter-trigram encoding. Baseline is the standard utterance-scope LSTM-LM. We see a large perplexity reduction of 17-21% by conditioning on session history words, with smaller incremental reductions from adding speaker change and overlap information.

The last two table rows show that some of the perplexity gain over the baseline is negated by the use of errorful recognition output for the conversation history. It does not make much difference whether the recognized word history is generated by just the subsystem being rescored ("single system", with 6% word error on SWB) or the full recognition system using multiple acoustic models ("full system", with about 5% word error rate on SWB and 10% on CH). Using recognition output as history, the perplexity degrades about 6% relative for SWB, and 11% on CH, relative to using the true word histories. Even with the more errorful recognition on CH, the session-based LM still gives a perplexity reduction of 14% relative to the baseline.

Table 2 presents recognition results, comparing baseline LSTM-LMs to the full session-based LSTM-LMs. Both the letter-trigram and one-word word encoding versions are reported. The different models may also be used jointly, using log-linear score combination in rescoring, shown in the third section of the table. We also tried iterating the session LM rescoring, after the recognized word histories were updated from the first rescoring pass (shown as "2nd iteration" in the table).

Results show that the session-based LM yields between 1% and 4% relative word error reduction for the two word encodings, and test sets. When the two word encoding types are combined by log-linear combination of model scores, the gain from session-based modeling is preserved. Iterating the session LM rescoring to improve the word histories did not give consistent gains.

Even though the session-based LSTM subsumes all the information used in the standard LSTM, there is an additional gain to be had from combining those two model types (last row in the table). Thus, the overall gain from adding the session-based models to the two baseline models is 3-5% relative word error reduction.

6 Conclusion and Future Work

We have proposed a simple generalization of utterance-level LSTM language models aimed at capturing conversational phenomena that operate across utterances and speakers, such as lexical entrainment, adjacency pairs, speech overlap, and topical coherence. To capture non-local conditioning information, the LSTM-LM is trained to read the entire sequence of utterances making up a conversation, along with side information encoding speaker changes and overlap of utterances. This is found to reduce perplexity by about 25%, most of which is retained when errorful recognition output is used to represent the word history in previous utterances. The session-based LM yields up to 5% relative reduction in word error when the utterance- and session-based LMs are combined.

It would be worthwhile to investigate which conversational phenomena are actually being exploited by the session LSTM model. The ease with which additional information can be input to the LSTM-LM also suggests encoding other conditioning information, such a more details about utterance timing, as well as semantic features that capture topical coherence.

References

Jerome R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42:93–108.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Studies in Fuzziness and Soft Computing*, volume 194, pages 137–186.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of HLT-NAACL 2003, Conference of the North American Chapter of the Association of Computational Linguistics*, volume 2, pages 7–9, Edmonton, Alberta, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Gang Ji and Jeffrey Bilmes. 2004. Multi-speaker language modeling. In *Proceedings of HLT-NAACL 2004, Conference of the North American Chapter of the Association of Computational Linguistics*, volume Short Papers, pages 133–136, Boston. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. Proceedings 3rd International Conference for Learning Representations, arXiv preprint arXiv:1703.02136.

Ivan Medennikov, Alexey Prudnikov, and Alexander Zatvornitskiy. 2016. Improving English conversational telephone speech recognition. In *Proc. Interspeech*, pages 2–6.

Microsoft Research. 2016. The Microsoft Cognition Toolkit (CNTK). https://cntk.ai.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. Interspeech*, pages 1045–1048.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *Proc. Interspeech*, pages 901–904.

H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest semantics for the organization of turn-taking in conversation. *Language*, 50(4):696–735.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English conversational telephone speech recognition by humans and machines. In *Proc. Interspeech*, pages 132–136, Stockholm.

E. A. Schegloff. 1968. Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 2014. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Proc. Interspeech*, pages 1058–1062.

Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, volume 2, pages 1359–1362, Aalborg, Denmark.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Proc. Interspeech*, pages 194–197.

W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2018. The Microsoft 2017 conversational speech recognition system. In *Proc. IEEE ICASSP*, pages 5934–5938, Calgary, Alberta.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. Toward human parity in conversational speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423.

D. Yu et al. 2014. An introduction to computational networks and the Computational Network Toolkit. Technical Report MSR-TR-2014-112, Microsoft Research. Https://github.com/Microsoft/CNTK.