

Code-switched Language Models Using Dual RNNs and Same-Source Pretraining

Saurabh Garg*

Tanmay Parekh*

Preethi Jyothi

Indian Institute of Technology, Bombay
{saurabhgarg,tanmayb,pjyothi}@cse.iitb.ac.in

Abstract

This work focuses on building language models (LMs) for code-switched text. We propose two techniques that significantly improve these LMs: 1) A novel recurrent neural network unit with dual components that focus on each language in the code-switched text separately 2) Pretraining the LM using synthetic text from a generative model estimated using the training data. We demonstrate the effectiveness of our proposed techniques by reporting perplexities on a Mandarin-English task and derive significant reductions in perplexity.

1 Introduction

Code-switching is a widespread linguistic phenomenon among multilingual speakers that involves switching between two or more languages in the course of a single conversation or within a single sentence (Auer, 2013). Building speech and language technologies to handle code-switching has become a fairly active area of research and presents a number of interesting technical challenges (Çetinoglu et al., 2016). Language models for code-switched text is an important problem with implications to downstream applications such as speech recognition and machine translation of code-switched data. A natural choice for building such language models would be to use recurrent neural networks (RNNs) (Mikolov et al., 2010), which yield state-of-the-art language models in the case of monolingual text. In this work, we explore mechanisms that can significantly improve upon such a baseline when applied to code-switched text. Specifically, we develop two such mechanisms:

- We alter the structure of an RNN unit to include separate components that focus on each

language in code-switched text separately while coordinating with each other to retain contextual information across code-switch boundaries. Our new model is called a *Dual RNN Language Model* (D-RNNLM), described in Section 2.

- We propose using *same-source pretraining* – i.e., pretraining the model using data sampled from a generative model which is itself trained on the given training data – before training the model on the same training data (see Section 3). We find this to be a surprisingly effective strategy.

We study the improvements due to these techniques under various settings (e.g., with and without access to monolingual text in the candidate languages for pretraining). We use perplexity as a proxy to measure the quality of the language model, evaluated on code-switched text in English and Mandarin from the SEAME corpus. Both the proposed techniques are shown to yield *significant perplexity improvements (up to 13% relative)* over different baseline RNNLM models (trained with a number of additional resources). We also explore how to combine the two techniques effectively.

Related Work: Adel et al. (2013) was one of the first works to explore the use of RNNLMs for code-switched text. Many subsequent works explored the use of external sources to enhance code-switched LMs, including the use of part-of-speech (POS) tags, syntactic and semantic features (Yeh et al., 2010; Adel et al., 2014, 2015) and the use of machine translation systems to generate synthetic text (Vu et al., 2012). Prior work has also explored the use of interpolated LMs trained separately on monolingual texts (Bhuvana-giri and Kopparapu, 2010; Imseng et al., 2011; Li et al., 2011; Baheti et al., 2017). Linguistic constraints governing code-switching have also

*Joint first authors

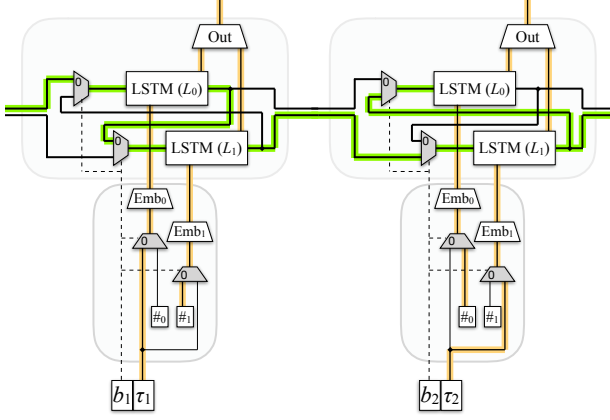


Figure 1: Illustration of the dual RNNLM (see the text for a detailed description). The highlighted left-to-right path (in green) indicates the flow of state information, when $b_1 = 0$ and $b_2 = 1$ (corresponding to token τ_1 belonging to language L_0 and τ_2 belonging to L_1). The highlighted bottom-to-top path (in orange) indicates the inputs and outputs.

been used as explicit priors to model when people switch from one language to another. Following this line of enquiry, (Chan et al., 2004) used grammar rules to model code-switching; (Li and Fung, 2013, 2014) incorporated syntactic constraints with the help of a code-switch boundary prediction model; (Pratapa et al., 2018) used a linguistically motivated theory to create grammatically consistent synthetic code-mixed text.

2 Dual RNN Language Models

Towards improving the modeling of code-switched text, we introduce Dual RNN Language Models (D-RNNLMs). The philosophy behind D-RNNLMs is that two different sets of neurons will be trained to (primarily) handle the two languages. (In prior work (Garg et al., 2018), we applied similar ideas to build dual N-gram based language models for code-switched text.)

As shown in Figure 1, the D-RNNLM consists of a “Dual LSTM cell” and an input encoding layer. The Dual LSTM cell, as the name indicates, has two long short-term memory (LSTM) cells within it. The two LSTM cells are designated to accept input tokens from the two languages L_0 and L_1 respectively, and produce an (unnormalized) output distribution over the tokens in the same language. When a Dual LSTM cell is invoked with an input token τ , the two cells will be invoked sequentially. The first (upstream) LSTM cell corresponds to the language that τ belongs to, and gets τ as its input. It passes on the resulting state to the downstream LSTM cell (which takes

a dummy token as input). The unnormalized outputs from the two cells are combined and passed through a soft-max operation to obtain a distribution over the union of the tokens in the two languages. Figure 1 shows a circuit representation of this configuration, using multiplexers (shaded units) controlled by a selection bit b_i such that the i^{th} token τ_i belongs to L_{b_i} .

The input encoding layer also uses multiplexers to direct the input token to the upstream LSTM cell. Two dummy tokens $\#_0$ and $\#_1$ are added to L_0 and L_1 respectively, to use as inputs to the downstream LSTM cell. The input tokens are encoded using an embedding layer of the network (one for each language), which is trained along with the rest of the network to minimize a cross-entropy loss function.

The state-update and output functions of the Dual LSTM cell can be formally described as follows. It takes as input (b, τ) where b is a bit and τ is an input token, as well as a state vector of the form (h_0, h_1) corresponding to the state vectors produced by its two constituent LSTMs. Below we denote the state-update and output functions of these two LSTMs as $H_b(\tau, h)$ and $O_b(\tau, h)$ (for $b = 0, 1$):

$$\begin{aligned}
 H((b, \tau), (h_0, h_1)) &= (h'_0, h'_1) \quad \text{where} \\
 (h'_0, h'_1) &= \begin{cases} (H_0(\tau, h_0), H_1(\#_1, h'_0)) & \text{if } b = 0 \\ (H_0(\#_0, h'_1), H_1(\tau, h_1)) & \text{if } b = 1 \end{cases} \\
 O((b, \tau), (h_0, h_1)) &= \text{softmax}(o_0, o_1) \quad \text{where} \\
 (o_0, o_1) &= \begin{cases} (O_0(\tau, h_0), O_1(\#_1, h'_0)) & \text{if } b = 0 \\ (O_0(\#_0, h'_1), O_1(\tau, h_1)) & \text{if } b = 1. \end{cases}
 \end{aligned}$$

Note that above, the inputs to the downstream LSTM functions H_{1-b} and O_{1-b} are expressed in terms of h'_b which is produced by the upstream LSTM.

3 Same-Source Pretraining

Building robust LMs for code-switched text is challenging due to the lack of availability of large amounts of training data. One solution is to artificially generate code-switched to augment the training data. We propose a variant of this approach – called same-source pretraining – in which the actual training data itself is used to train a generative model, and the data sampled from this model is used to pretrain the language model.

Same-source pretraining can leverage powerful training techniques for generative models to train a

| | Train | Dev | Test |
|-------------------|---------|---------|---------|
| # Utterances | 74,927 | 9,301 | 9,552 |
| # Tokens | 977,751 | 131,230 | 114,546 |
| # English Tokens | 316,726 | 30,154 | 50,537 |
| # Mandarin Tokens | 661,025 | 101,076 | 64,009 |

Table 1: Statistics of data splits derived from SEAME.

language model. We note that the generative models by themselves are typically trained to minimize a different objective function (e.g., a discrimination loss) and need not perform well as language models.*

Our default choice of generative model will be an RNN (but see the end of this paragraph). To complete the specification of same-source pre-training, we need to specify how it is trained from the given data. Neural language models trained using the maximum likelihood training paradigm tend to suffer from the *exposure bias* problem during inference when the model generates a text sequence by conditioning on previous tokens that may never have appeared during training. Scheduled sampling (Bengio et al., 2015) can help bridge this gap between the training and inference stages by using model predictions to synthesize prefixes of text that are used during training, rather than using the actual text tokens. A more promising alternative to generate text sequences was recently proposed by Yu et al. (2017) where sequence generation is modeled in a generative adversarial network (GAN) based framework. This model – referred to as “SeqGAN” – consists of a generator RNN and a discriminator network trained as a binary classifier to distinguish between real and generated sequences. The main innovation of SeqGAN is to train the generative model using policy gradients (inspired by reinforcement learning) and use the discriminator to determine the reward function. We experimented with using both naïve and scheduled sampling based training; using SeqGAN was a consistently better choice (5 points or less in terms of test perplexities) compared to these two sampling methods. As such, we use SeqGAN as our training method for the generator. We also experiment with replacing the RNN with a Dual RNN as the generator in the SeqGAN training and observe small but consistent reductions in perplexity.

*In our experiments, we found the perplexity measures for the generative models to be an *order of magnitude* larger than that of the LMs we construct.

4 Experiments and Results

Dataset Preparation: For our experiments, we use code-switched text from the SEAME corpus (Lyu et al., 2010) which contains conversational speech in Mandarin and English. Since there is no standardized task based on this corpus, we construct our own training, development and test sets using a random 80-10-10 split. Table 1 shows more details about our data sets. (Speakers were kept disjoint across these datasets.)

Evaluation Metric: We use token-level perplexity as the evaluation metric where tokens are words in English and characters in Mandarin. The SEAME corpus provides word boundaries for Mandarin text. However, we used Mandarin characters as individual tokens since a large proportion of Mandarin words appeared very sparsely in the data. Using Mandarin characters as tokens helped alleviate this issue of data sparsity; also, applications using Mandarin text are typically evaluated at the character level and do not rely on having word boundary markers (Vu et al., 2012).

Outline of Experiments: Section 4.1 will explore the benefits of both our proposed techniques – (1) using D-RNNLMs and (2) using text generated from SeqGAN for pretraining – in isolation and in combination. Section 4.2 will introduce two additional resources (1) monolingual text for pre-training and (2) a set of syntactic features used as additional input to the RNNLMs that further improve baseline perplexities. We show that our proposed techniques continue to outperform the baselines albeit with a smaller margin. All these perplexity results have been summarized in Table 2.

4.1 Improvements Over the Baseline

This section focuses only on the numbers listed in the first two columns of Table 2. The Baseline model is a 1-layer LSTM LM with 512 hidden nodes, input and output embedding dimensionality of 512, trained using SGD with an initial learning rate of 1.0 (decayed exponentially after 80 epochs at a rate of 0.98 till 100 epochs) The development and test set perplexities using the baseline are 89.60 and 74.87, respectively.

The D-RNNLM is a 1-layer language model with each LSTM unit having 512 hidden nodes. The training paradigm is similar to the above-mentioned setting for the baseline model.[†] We see

[†]D-RNNLMs have a few additional parameters. How-

| | w/o syntactic features | | | | with syntactic features | | | |
|---------------------|------------------------|--------------|-----------------|--------------|-------------------------|--------------|-----------------|--------------|
| | w/o mono. data | | with mono. data | | w/o mono. data | | with mono. data | |
| | Dev | Test | Dev | Test | Dev | Test | Dev | Test |
| Baseline | 89.60 | 74.87 | 74.06 | 61.66 | 81.87 | 68.23 | 71.04 | 59.00 |
| D-RNNLM | 88.68 | 72.29 | 72.41 | 60.73 | 81.01 | 66.26 | 70.83 | 59.04 |
| With RNNLM SeqGAN | 79.16 | 65.96 | 72.51 | 60.56 | 77.30 | 63.75 | 68.43 | 55.71 |
| With D-RNNLM SeqGAN | 78.63 | 65.41 | 72.33 | 60.30 | 77.19 | 63.63 | 67.79 | 55.60 |

Table 2: Development set and test set perplexities using RNNLMs and D-RNNLMs with various pretraining strategies.

consistent improvements in test perplexity when comparing a D-RNNLM with an RNNLM (i.e. 74.87 drops to 72.29).[‡]

Next, we use text generated from a SeqGAN model to pretrain the RNNLM.[§] We use our best trained RNNLM baseline as the generator within SeqGAN. We sample 157,440 sentences (with a fixed sentence length of 20) from the SeqGAN model; this is thrice the amount of code-switched training data. We first pretrain the baseline RNNLM with this sampled text, before training it again on the code-switched text. This gives significant reductions in test perplexity, bringing it down to 65.96 (from 74.87).

Finally, we combine both our proposed techniques by replacing the generator with our best-trained D-RNNLM within SeqGAN. Although there are other ways of combining both our proposed techniques, e.g. pretraining a D-RNNLM using data sampled from an RNNLM SeqGAN, we found this method of combination to be most effective. We see modest but consistent improvements with D-RNNLM SeqGAN over RNNLM SeqGAN in Table 2, further validating the utility of D-RNNLMs.

4.2 Using Additional Resources

We employed two additional resources to further improve our baseline models. First, we used monolingual text in the candidate languages to pretrain the RNNLM and D-RNNLM models. We used transcripts from the Switchboard corpus[¶] for English; AIShell^{||} and THCHS30^{**} corpora for

ever, increasing the capacity of an RNNLM to exactly match this number makes its test perplexity worse; RNNLM with 720 hidden units gives a development set perplexity of 91.44 and 1024 hidden units makes it 91.46.

[‡]Since D-RNNLMs use language ID information, we also trained a baseline RNNLM with language ID features; this did not help reduce the baseline test perplexities. In future work, we will explore alternate LSTM-based models that incorporate language ID information (Chandu et al., 2018)

[§]To implement SeqGAN, we use code from <https://github.com/LantaoYu/SeqGAN>.

[¶]<http://www.openslr.org/5/>

^{||}<http://www.openslr.org/33/>

^{**}<http://www.openslr.org/18/>

Mandarin monolingual text data. This resulted in a total of ≈ 3.1 million English tokens and ≈ 2.5 million Mandarin tokens. Second, we used an additional set of input features to the RNNLMs and D-RNNLMs that were found to be useful for code-switching in prior work (Adel et al., 2014). The feature set included part-of-speech (POS) tag features and Brown word clusters (Brown et al., 1992), along with a language ID feature. We extracted POS tags using the Stanford POS-tagger^{††} and we clustered the words into 70 classes using the unsupervised clustering algorithm by Brown et al. (1992) to get Brown cluster features.

The last six columns in Table 2 show the utility of using either one of these resources or both of them together (shown in the last two columns). The perplexity reductions are largest (compared to the numbers in the first two columns) when combining both these resources together. Interestingly, all the trends we observed in Section 4.1 still hold. D-RNNLMs still consistently perform better than their RNNLM counterparts and we obtain the best overall results using D-RNNLM SeqGAN.

5 Discussion and Analysis

| | Eng-Eng | Eng-Man | Man-Eng | Man-Man |
|----------------|---------|---------|---------|---------|
| RNNLM | 133.18 | 157.18 | 2617.28 | 34.98 |
| D-RNNLM | 140.37 | 151.38 | 2452.16 | 32.89 |
| Mono RNNLM | 101.61 | 181.28 | 2510.48 | 30.00 |
| Mono D-RNNLM | 101.66 | 156.44 | 2442.81 | 29.64 |
| RNNLM SeqGAN | 120.28 | 154.44 | 2739.85 | 30.40 |
| D-RNNLM SeqGAN | 120.26 | 149.68 | 2450.85 | 30.60 |

Table 3: Decomposed perplexities on the development set on all four types of tokens from various models.

Table 3 shows how the perplexities on the development set from six of our prominent models decompose into the perplexities contributed by English tokens preceded by English tokens (Eng-Eng), Eng-Man, Man-Eng and Man-Man tokens. This analysis reveals a number of interesting observations. 1) The D-RNNLM mainly improves over the baseline on the “switching tokens”, Eng-

^{††}<https://nlp.stanford.edu/software/tagger.shtml>

Man and Man-Eng. 2) The RNNLM with monolingual data improves most over the baseline on “the monolingual tokens”, Eng-Eng and Man-Man, but suffers on the Eng-Man tokens. The D-RNNLM with monolingual data does as well as the baseline on the Eng-Man tokens and performs better than “Mono RNNLM” on all other tokens. 3) RNNLM SeqGAN suffers on the Man-Eng tokens, but helps on the rest; in contrast, D-RNNLM SeqGAN helps on *all tokens* when compared with the baseline.

| | SeqGAN-RNNLM | SeqGAN-DLM |
|----------|--------------|------------|
| Bigram | 25.57 | 31.33 |
| Trigram | 75.88 | 83.86 |
| Quadgram | 137.98 | 145.71 |

Table 4: Percentage of new n-grams generated.

As an additional measure of the quality of text generated by RNNLM SeqGAN and D-RNNLM SeqGAN, in Table 4, we measure the diversity in the generated text by looking at the increase in the number of unique n-grams with respect to the SEAME training text. D-RNNLM SeqGAN is clearly better at generating text with larger diversity, which could be positively correlated with the perplexity improvements shown in Table 2.

While we do not claim same-source pretraining may be an effective strategy in general, we show it is useful in low training-data scenarios. Even with only $\frac{1}{16}$ th of the original SEAME training data used for same-source pretraining, development and test perplexities are reduced to 84.45 and 70.59, respectively (compared to 79.16 and 65.96 using the entire training data).

6 Conclusion

D-RNNLMs and same-source pretraining provide significant perplexity reductions for code-switched LMs. These techniques may be of more general interest. Leveraging generative models to train LMs is potentially applicable beyond code-switching; D-RNNLMs could be generalized beyond LMs, e.g. speaker diarization. We leave these for future work to explore.

7 Acknowledgments

The authors thank the anonymous reviewers for their helpful comments and suggestions. The last author gratefully acknowledges financial support from Microsoft Research India (MSRI) for this project.

References

- Heike Adel, Dominic Telaar, Ngoc Thang Vu, Katrin Kirchhoff, and Tanja Schultz. 2014. Combining recurrent neural networks and factored language models during decoding of code-switching speech. In *Proceedings of Interspeech*.
- Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar, and Tanja Schultz. 2015. Syntactic and semantic features for code-switching factored language models. *Proceedings of IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):431–440.
- Heike Adel, Ngoc Thang Vu, Franziska Kraus, Tim Schlippe, Haizhou Li, and Tanja Schultz. 2013. Recurrent neural network language modeling for code switching conversational speech. In *Proceedings of ICASSP*, pages 8411–8415. IEEE.
- Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.
- Ashutosh Baheti, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. In *Proceedings of ICON*, pages 65–74.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of NIPS*, pages 1171–1179.
- K Bhuvanagiri and Sunil Kopparapu. 2010. An approach to mixed language automatic speech recognition. *Proceedings of Oriental COCOSA, Kathmandu, Nepal*.
- Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.
- Özlem Çetinoglu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. *Proceedings of EMNLP*, page 1.
- Joyce YC Chan, PC Ching, Tan Lee, and Helen M Meng. 2004. Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Proceedings of International Symposium on Chinese Spoken Language Processing*, pages 293–296. IEEE.
- Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W Black. 2018. Language informed modeling of code-switched text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 92–97.
- Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Dual language models for code mixed speech recognition. In *Proceedings of Interspeech*.

- David Imseng, Hervé Bourlard, Mathew Magimai Doss, and John Dines. 2011. Language dependent universal phoneme posterior estimation for mixed language speech recognition. In *Proceedings of ICASSP*, pages 5012–5015.
- Ying Li and Pascale Fung. 2013. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *Proceedings of ICASSP*, pages 7368–7372. IEEE.
- Ying Li and Pascale Fung. 2014. Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of EMNLP*.
- Ying Li, Pascale Fung, Ping Xu, and Yi Liu. 2011. Asymmetric acoustic modeling of mixed language speech. In *Proceedings of ICASSP*, pages 5004–5007.
- Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. 2010. An analysis of a Mandarin-English code-switching speech corpus: SEAME. *Proceedings of Age*, 21:25–8.
- Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*.
- Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of ACL*, volume 1, pages 1543–1553.
- Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for Mandarin-English code-switch conversational speech. In *Proceedings of ICASSP*, pages 4889–4892. IEEE.
- Ching Feng Yeh, Chao Yu Huang, Liang Che Sun, and Lin Shan Lee. 2010. An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling. In *Proceedings of Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 214–219. IEEE.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of AAAI*, pages 2852–2858.