LSTM Neural Reordering Model for Statistical Machine Translation

Yiming Cui, Shijin Wang, Jianfeng Li

iFLYTEK Research

June 14, 2016

OUTLINE

- Lexicalized Reordering Model
- LSTM Neural Reordering Model
- Experiments & Analyses
- Related Work
- Conclusion & Future Work
- References

LEXICALIZED RM

- Lexicalized Reordering Model
  - The most widely used RM
  - Given source and target sentence **f,e** and phrase alignment **a**

$$p(\boldsymbol{o}|\boldsymbol{e},\boldsymbol{f}) = \prod_{i=1}^{n} P(o_i|e_i, f_{a_i}, a_{i-1}, a_i)$$

LEXICALIZED RM

- Lexicalized Reordering Model
  - orientation type **o**: LR, MSD, MSLR
  - Take MSD type for e.g., it can be defined as

$$o_i = \begin{cases} M, if\ a_i - a_{i-1} = 1 \\ S, if\ a_i - a_{i-1} = -1 \\ D, if\ |a_i - a_{i-1}| \neq 1 \end{cases}$$

Lexicalized RM

- Lexicalized Reordering Model
  - Some researcher also suggested that by including both current and previous phrase pairs into condition, can improve accuracy (Li et al., 2014)

$$P(\mathbf{o}|\mathbf{f},\mathbf{e},\mathbf{a}) \approx \prod_{i=1}^{n} P(o_i|\tilde{f}_{a_i},\tilde{e}_i,a_{i-1},a_i)$$

$$\Downarrow$$

$$P(\mathbf{o}|\mathbf{f},\mathbf{e},\mathbf{a}) \approx \prod_{i=1}^{n} P(o_i|\tilde{f}_{a_i},\tilde{e}_i,\tilde{f}_{a_{i-1}},\tilde{e}_{i-1},a_{i-1},a_i)$$

LSTM Neural RM

- Why RNN?
  - RNNs are capable to learn sequential problems
  - It is natural to use RNNs to include much more history to predict next word's orientation (reordering)
  - Further by utilizing LSTM, RNNs are able to capture long-time dependency, and solve "Gradient Vanishing" problem (Bengio, 1997)

LSTM Neural RM

- Training data processing
  - Given source and target sentence pair and alignment

(1) If current target word is one-to-one alignment, then we can directly induce its orientations (left or right).

(2) If current source/target word is one-to-many alignment, then we judge its orientation by considering its first aligned target/source word, and the other aligned target/source words are annotated as “<follow>” reordering type, which means this word pair inherent orientation of previous word pair.

(3) If current source/target word is not aligned to any target/source words, we introduce a “<NULL>” token in the opposite side, and annotate this word pair as “<follow>” reordering type.

LSTM Neural RM

- Training Data Processing: Example

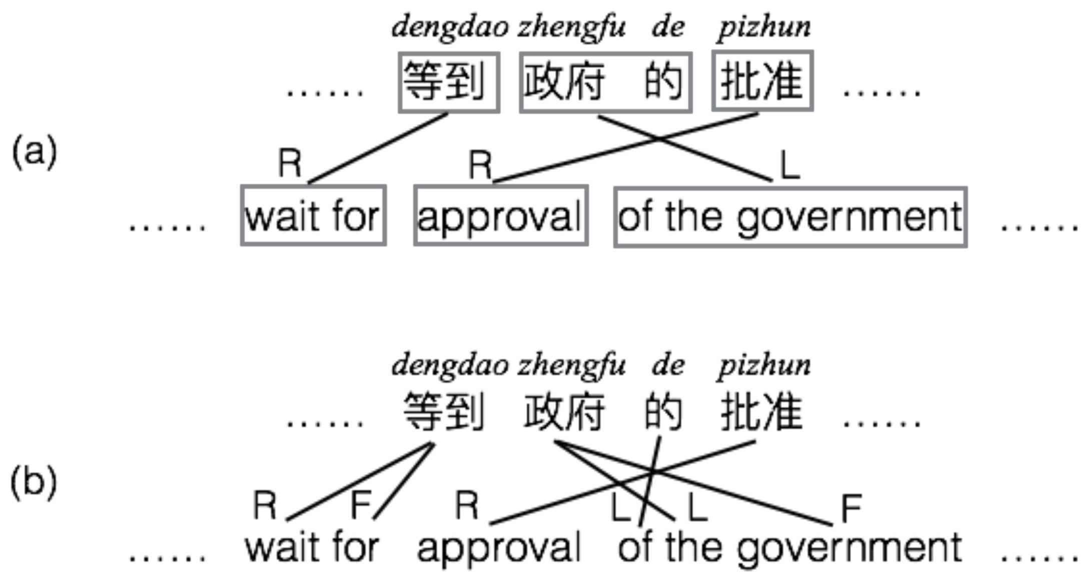


Figure 1: Illustration of data processing. (a) Original reordering (omit the alignment inside phrase); (b) processed reordering, "R"-right, "L"-left, "F"-follow.

LSTM NEURAL RM

- History Extended Reordering Model

$$P(\mathbf{o}|\mathbf{f},\mathbf{e},\mathbf{a}) \approx \prod_{i=1}^{n} P(o_i|\tilde{f}_{a_i}, \tilde{e}_i, a_{i-1}, a_i)$$

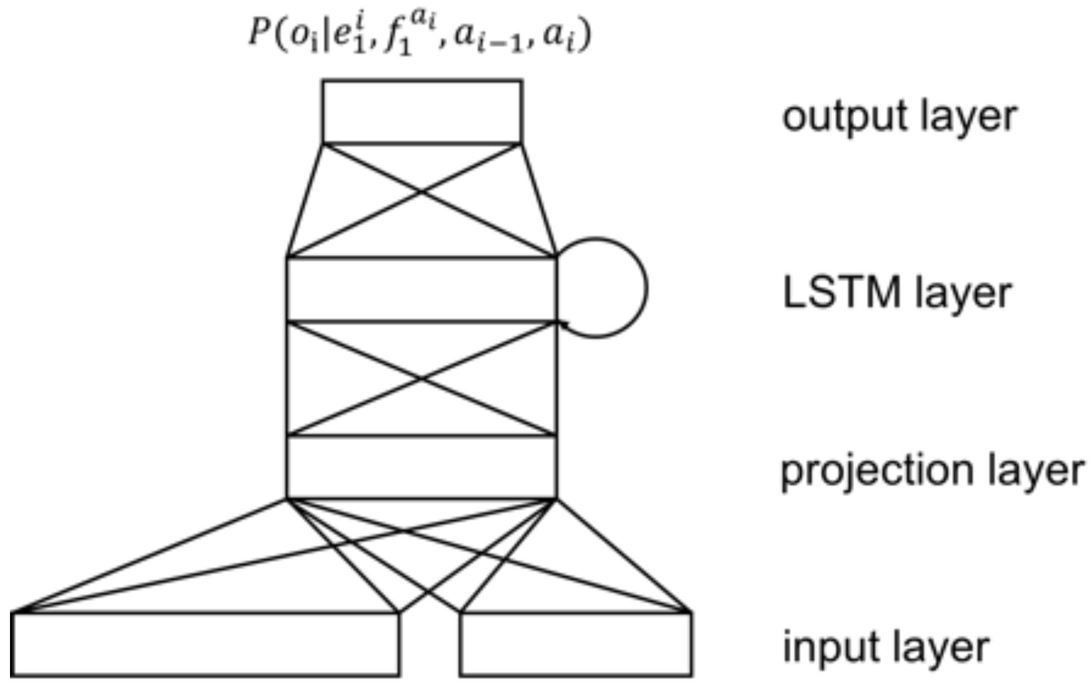
$$P(\mathbf{o}|\mathbf{f},\mathbf{e},\mathbf{a}) \approx \prod_{i=1}^{n} P(o_i|\tilde{f}_{a_i}, \tilde{e}_i, \tilde{f}_{a_{i-1}}, \tilde{e}_{i-1}, a_{i-1}, a_i)$$

Proposed model

$$p(\boldsymbol{o}|\boldsymbol{e},\boldsymbol{f}) = \prod_{i=1}^{n} P(o_i|e_1^i, f_1^{a_i}, a_{i-1}, a_i)$$

LSTM NEURAL RM

- LSTM NRM Architecture



$$y_i = W_1 * f_{a_i} + W_2 * e_i$$
$$z_i = LSTM(y_i, W_3, y_1^{i-1})$$
$$P\left(o_i \middle| e_1^i, f_1^{a_i}, a_{i-1}, a_i\right) = softm(W_4 * z_i)$$

Experiment

- Setups
  - NIST OpenMT12 ZH-EN and AR-EN Task
  - Apply RNNRM into N-best rescoring step
  - Results are average with 5 runs (Clark et al., 2011)
  - Neural params: hidden units 100, SGD(alpha=0.01), source-vocab 100k, target-vocab 50k

EXPERIMENT

- Results on different orientation types
- All results are significantly better than each baseline, using paired bootstrap resampling method (Koehn, 2004)

System	Dev	Test1	Test2
Baseline	43.87	39.84	42.05
+LR	44.43	40.53	**42.84**
+MSD	44.29	40.41	42.62
+MSLR	**44.52**	**40.59**	42.78

Table 2: LSTM reordering model with different orientation types for Arabic-English system.

System	Dev	Test1	Test2
Baseline	27.18	26.17	24.04
+LR	**27.90**	26.58	**24.59**
+MSD	27.49	26.51	24.39
+MSLR	27.82	**26.78**	24.53

Table 3: LSTM reordering model with different orientation types for Chinese-English system.

EXPERIMENT

- Results on different reordering baselines

Ar-En System	Dev	Test1	Test2
Baseline_wbe	43.87	39.84	42.05
+NRM_MSLR	44.52	40.59	42.78
Baseline_phr	44.11	40.09	42.21
+NRM_MSLR	44.52	40.73	42.89
Baseline_hier	44.30	40.23	42.38
+NRM_MSLR	44.61	40.82	42.86

Zh-En System	Dev	Test1	Test2
Baseline_wbe	27.18	26.17	24.04
+NRM_MSLR	27.90	26.58	24.70
Baseline_phr	27.33	26.05	24.13
+NRM_MSLR	27.86	26.46	24.73
Baseline_hier	27.56	26.29	24.38
+NRM_MSLR	28.02	26.49	24.67

Related Work

- Neural network based approach has been widely applied into SMT field
  - LM: NNLM(Bengio et al., 2003), RNNLM(Mikolov et al., 2011)
  - TM: NNJM(Devlin et al., 2014), RNNTM(Sundermeyer et al., 2014)
  - RM: RAE classification method (Li et al., 2014)

CONCLUSION & FUTURE WORK

- Conclusion
  - propose a purely lexicalized neural reordering model
  - support different orientation types: LR/MSD/MSLR
  - Easily integrate into rescoring & outperform baseline systems
- Future Work
  - Dissolve much more ambiguities and improve reordering accuracy by introducing phrase-based
  - Apply NRM into NMT

REFERENCES

- Y. Bengio, P. Simard, and P. Frasconi. 1994. Learn- ing long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166.
- Yoshua Bengio, Holger Schwenk, Jean Sbastien Sencal, Frderic Morin, and Jean Luc Gauvain. 2003. A neu- ral probabilistic language model. Journal of Machine Learning Research, 3(6):1137–1155.
- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In Pro- ceedings of the 2012 Conference of the North Ameri- can Chapter of the Association for Computational Lin- guistics: Human Language Technologies, pages 427– 436, Montre ́al, Canada, June. Association for Compu- tational Linguistics.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statisti- cal machine translation: Controlling for optimizer in- stability. In Proceedings of the 49th Annual Meet- ing of the Association for Computational Linguistics: Human Language Technologies, pages 176–181, Port- land, Oregon, USA, June. Association for Computa- tional Linguistics.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for sta- tistical machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Compu- tational Linguistics (Volume 1: Long Papers), pages 1370–1380, Baltimore, Maryland, June. Association for Computational Linguistics.

REFERENCES

- Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 848–856, Honolulu, Hawaii, October. Association for Computational Linguistics.
- A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm net- works. In Proceedings in 2005 IEEE International Joint Conference on Neural Networks, pages 2047– 2052 vol. 4.Alex Graves. 1997. Long short-term memory. Neural Computation, 9(8):1735–1780.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2004. Statistical phrase-based translation. In Con- ference of the North American Chapter of the Asso- ciation for Computational Linguistics on Human Lan- guage Technology-volume, pages 127–133.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Con- stantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceed- ings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Pro- ceedings of the Demo and Poster Sessions, pages 177– 180, Prague, Czech Republic, June. Association for Computational Linguistics.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, Proceedings of EMNLP 2004, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

REFERENCES

- Peng Li, Yang Liu, and Maosong Sun. 2013. Recursive autoencoders for ITG-based translation. In Proceed- ings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 567–577, Seat- tle, Washington, USA, October. Association for Com- putational Linguistics.
- Peng Li, Yang Liu, Maosong Sun, Tatsuya Izuha, and Dakun Zhang. 2014. A neural reordering model for phrase-based translation. In Proceedings of COLING 2014, the 25th International Conference on Compu- tational Linguistics: Technical Papers, pages 1897– 1907, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.
- T. Mikolov, S. Kombrink, L. Burget, and J. H. Cernocky. 2011. Extensions of recurrent neural network lan- guage model. In IEEE International Conference on Acoustics, Speech Signal Processing, pages 5528– 5531.
- Franz Josef Och and Hermann Ney. 2000. A compari- son of alignment models for statistical machine trans- lation. In Proceedings of the 18th conference on Com- putational linguistics - Volume 2, pages 1086–1090.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. Bleu: a method for automatic eval- uation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylva- nia, USA, July. Association for Computational Lin- guistics.
- Andreas Stolcke. 2002. Srilm — an extensible language modeling toolkit. In Proceedings of the 7th Inter-national Conference on Spoken Language Processing (ICSLP 2002), pages 901–904.

REFERENCES

- Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In Proceed- ings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 14–25, Doha, Qatar, October. Association for Computational Linguistics.
- Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, HLT- NAACL 2004: Short Papers, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
- Ashish Vaswani, Liang Huang, and David Chiang. 2012. Smaller alignment models for better translations: Un- supervised word alignment with the l0-norm. In Pro- ceedings of the 50th Annual Meeting of the Associa- tion for Computational Linguistics (Volume 1: Long Papers), pages 311–319, Jeju Island, Korea, July. As- sociation for Computational Linguistics.

Thank You !