BILINGUAL RECURRENT NEURAL NETWORKS FOR IMPROVED STATISTICAL MACHINE TRANSLATION

Bing Zhao[*]

LinkedIn Corporation
2029 Stierlin Ct, Mountain View,
CA 94043, USA

Yik-Cheung Tam

SRI International
333 Ravenswood Ave, Menlo Park,
CA 94025, USA

ABSTRACT

Recurrent Neural Networks (RNN) have been successfully applied for improved speech recognition and statistical machine translation (SMT) for N-best list re-ranking. In SMT, we investigate using bilingual word-aligned sentences to train a bilingual recurrent neural network model. We employ a bag-of-word representation of a source sentence as additional input features in model training. Experimental results show that our proposed approach performs consistently better than recurrent neural network language model trained only on target-side text in terms of machine translation performance. We also investigate other input representation of a source sentence based on latent semantic analysis.

Index Terms— Bilingual recurrent neural network model, statistical machine translation

1. INTRODUCTION

Recurrent neural networks (RNN) have shown tremendous progress in improving many statistical NLP tasks in which language models (LM) play a key role [1, 2, 3, 4, 5, 6, 7, 8, 9], together with feed-forward continuous neural network for language modeling [10, 11], translation modeling [12, 13, 14], and the recent joint modeling of language and translation [15] that has yielded very impressive improvement in machine translation performance.

With the recursive nature, recurrent neural network language models (RNNLM) captures complex long-distance history across sentence boundaries. For example, the word "weapon" in the history of several sentences earlier, may increase the probability for the word "arms" in a dialogue at a checkpoint station; similarly the word "bank" in a dialogue history may increase the likelihood of the correct translation of the word "safe" (as in a secured deposit box).

In this paper, we focus on developing a bilingual recurrent neural network (bRNN) for statistical machine translation, leveraging information from source and target languages in a joint manner [16]. We first use a bag-of-word (BOW) representation for a source sentence as an additional sparse input for RNN. From a graphical modeling perspective, the previous word of a target language e_{j-1}, the previous hidden activation vector h_{j-1}, and the BOW vector of a source language F are the predictors for the current hidden activation vector h_j that predicts e_j. These variables form a V-shaped graphical structure. During bilingual RNN training, since e_j is observed, these predictors will compete with each other resulting in the so-called "explain-away" effect [17]. The BOW vector F may have a diluting effect on the influence of the previous target word e_{j-1} in predicting the next target word e_j since the word translation effect can be strong. Therefore, when predicting each target word e_j, we propose a "less-one" strategy to intentionally skip the aligned source words f_{a_j} and $f_{a_{j-1}}$ when creating the BOW vector F where a_j denotes a set of aligned source words of a target word e_j. Thus, we encourage the bilingual RNN training to focus on complementary source contextual information with respect to the previous target word.

The paper is organized as follows: In section 2, we present bilingual recurrent neural network and the variants; we describe experiments exploring different configurations for our proposed method in section 3. We provide discussions and conclusions in section 4.

2. BILINGUAL RECURRENT NEURAL NETWORK

Recurrent neural network language models trained on monolingual corpora have shown success in improving machine translation performance. One natural question is whether parallel corpora would be useful for learning a bilingual RNNLM to further improve machine translation performance. Intuitively, RNNLM trained on parallel corpora has a stronger word prediction power than the one trained on monolingual corpora because word prediction can be conditioned on the complete source sentence in addition to previous target word and hidden state vector. Given a set of parallel sentences and corresponding word alignment $\{F, E, A\}$ where $F = f_1...f_i...f_I$, $E = e_1...e_j...e_J$, and $A = a_1...a_j...a_J$ denote a source sentence, target sentence, and a word align-

[*] The author performed the work while at SRI International.

978-1-4799-7129-9/14/$31.00 ©2014 IEEE

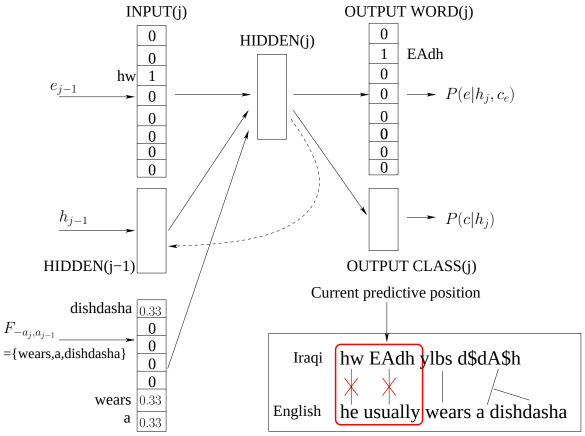

Fig. 1. Bag-of-Word approach with the "less-one" strategy: A bilingual recurrent neural network language model with an additional input for a source sentence F. The target language is Iraqi Arabic in the Buckwalter representation, and the source language is English. The alignment for the sentence pair is given, and the corresponding activated predictor is marked for predicting the current Iraqi token "EAdh". The "less-one" strategy is applied to intentionally remove the effect from the current and the previous aligned source words "usually" and "he", denoted as $F_{-a_j,a_{j-1}}$. With normalization, the source predictor activation sums to unity.

ment mapping a target word e_j to a set of word positions on the source sentence F, a generic bilingual RNNLM has the form $p(e_j|e_{j-1}, h_{j-1}, F, A)$. One way to exploit F is to introduce an additional input layer for F. [16] used the form $p(e_j|e_{j-1}, h_{j-1}, F)$ ignoring the word alignment. Then a dense feature vector is extracted from latent semantic analysis using a window of source words. Compared to RNNLM, their approach has shown drastic reduction in word perplexity, which is expected since "future" source words are given for predicting a target word. However, only a marginal gain is observed in machine translation performance compared to RNNLM (25.6 to 25.8 in BLEU). Their results would raise a question whether the source context F has minimal effect, or there exists another way to exploit F in a more fruitful manner.

2.1. Bag-of-Word representation

Our goal in bilingual RNN (bRNN) is to model complementary information that has not been well modeled by conventional models such as the IBM word/phrase translation models. First, we use a "bag-of-word" (BOW) approach on a source sentence to construct a sparse input feature vector. In the "bag-of-word" model, word frequency is not taken into account in order to minimize the effect of frequent function words like "the". [18] has also used "bag-of-word" representation for language understanding in a monolingual setting. To minimize direct modeling of source-to-target word translations that has been covered by IBM-1 model, i.e. $p(e|f)$, we propose the "less-one" approach, in which we deactivate the source words that are direct translations to the current and previous target word according to word alignment. The reasons are two-folded: (1) these translation pairs may be well captured in IBM translation models that are used as features in the log-linear model; (2) these aligned source words can dominate other predictors, i.e. explain-away the previous target word predictor that is indeed informative as in RNNLM. Figure 1 summarizes our "less-one" approach. When predicting a target word e_j, we remove source words according to the alignment positions a_j and a_{j-1} when a "bag-of-word" feature vector is constructed per target word prediction. With aligned source words deleted, we enforce bilingual RNNLM to learn complementary information such as cross-lingual word triggers from the source word context that may not be well covered by conventional IBM translation models. Putting these altogether, our bilingual RNNLM has the form $p(e_j|e_{j-1}, h_{j-1}, F_{-a_j,a_{j-1}})$ where $F_{-a_j,a_{j-1}}$ denotes removing source words according to word alignment variable a_j and a_{j-1} at the current target position j. When the current target word is unaligned, a_j is empty, and thus we use the whole source sentence in our model. In addition to the proposed "less-one" strategy, we perform the following aspects for bilingual RNNLM:

- To avoid dominating with the noisy predictors from the source BOW vector, we normalize the activation of each source word to be $1/L$ where L is the length of BOW so that the summation of source activation is unity.

- Use a well-trained RNNLM to partially initialize the weights of the bilingual RNN since they have overlapping network structure. By doing so, we hope to have a good starting point for learning the extra weights for BOW.

- When predicting the end-of-sentence symbol $< /s >$, we use a zero BOW vector for input. This essentially switches back to RNNLM since at this point, RNN has seen all word information from source and target languages already.

- Set the number of hidden unit of 600. Although this will increase the amount of training time significantly, using a smaller hidden unit of 200 has shown no benefit compared to RNNLM in our experiments.

2.2. Latent semantic analysis representation

To serve as a bilingual RNN baseline, we apply latent semantic analysis (LSA) to project an input source sentence into a low-dimension topic space represented by the topic posterior $p(k|F)$ where k denotes the latent topic index. In this work, we use latent Dirichlet-Tree allocation [19] with a tree-based Dirichlet prior for topic modeling. We use 100 topics throughout the reported experiments.

3. EXPERIMENTS

Our translation engine was built on data from the DARPA TRANSTAC program, a speech-to-speech translation initiative targeting tactical military communication. The source language was conversational English, and target language was Iraqi Arabic. This MT direction is more challenging, in that one needs to maintain valid morphology order in the MT output, and data scarcity is more prominent in Iraqi Arabic for LM training. We had $760K$ parallel sentence pairs as training data and 6985 sentence pairs for learning the log-linear weights for dense and sparse features. The tuning set had a single reference, and all test sets had 4 references. We filtered the tuning set by skipping short dialogues containing less than three sentences/turns as many of them were like simple sentences "thank you", or "you are welcome". Details are shown in Table 1.

Table 1. Data

Data	Sentences	Source words
Train	760200	7207779
Tune	6985	64193
Test1	567	6855
Test2	655	10652
Test3	617	9203

We applied a word segmenter on the Iraqi Arabic text to break a word into affixes. All models were built using the segmented data, and translations were post-processed into word forms for BLEU score computation. In our Hiero SMT baseline, we incorporated 12 dense features for each bilingual stochastic context-free grammar (SCFG) rule after the Hiero grammar in [20], including: IBM Model-1 scores in both source-to-target and target-to-source directions, relative frequencies in both directions, count of phrases, count of Hiero rules, number of source content words aligned to target spontaneous words, number of target spontaneous words aligned to source content words, three binned frequencies, and the number of unaligned source words. We further computed lexical-pairs seen in a dictionary, affix sequences/ngrams, fertility for each word in the SCFG rule, and additional spontaneous/content words mismatch as sparse features. In total, we had $368,524$ sparse features. Optimization methods such as MIRA [21] or PRO [22] were employed that can optimize millions of sparse features.

Our training data were marked with document boundaries, with $10,202$ naturally developed dialogues; each dialogue on average had 7 turns/sentences per speaker; each sentence had a time marker, so that we can sort the training data by turn sequences. Our tuning and test sets were marked with time information and dialogue boundaries. The sequential order of the training data was required to model discourse information in RNNLM training.

The training data were aligned using the grow-diag-final option with GIZA word alignment in both directions; then the aligned sentence pairs sorted with the sequential order were fed into bilingual RNN training, while the target side was fed into RNNLM training. 10% of the training data was kept for cross validation on word perplexity to ensure bilingual RNN training was on the right track, although word perplexity had little correlation towards final translation performance. Our bilingual RNN was implemented based on the RNNLM toolkit [4]. We employed 100 output classes and used back-propagation through time using the flag "-bptt 4 -bptt-block 10". Initial learning rate was chosen as 0.1, with the learning rate started to halve when the reduction in cross validation perplexity was small. We compared different setups using BLEU [23].

3.1. Reranking results

We applied our baseline translation engine to generate up to 2000 N-best hypotheses per source sentence. The combined weighted score was associated with each hypothesis so that the score was further combined with a score from RNNLM or bilingual RNN for reranking:

$$\text{score(rerank)} \quad = \quad \text{score(base)} + \lambda_{bRNN} \cdot f_{bRNN}(F, E)$$

where $f_{bRNN}(F, E)$ denotes the bilingual RNN score for a sentence pair F and E. λ_{bRNN} is optimized using a simple grid search.

Table 2 shows N-best reranking performance on BLEU using RNNLM and a variety of bilingual RNN models. Comparing with the SMT baseline with sparse features, RNNLM yielded 0.4 BLEU improvement across 3 test sets, and 1 BLEU improvement on test3. However, RNNLM had no improvement on the other test sets. Bilingual RNN using LSA to encode a source sentence failed to yield better performance than RNNLM. This results agreed with the observation in [16]. Bilingual RNN with BOW representation performed better than RNNLM by 0.1 BLEU overall. With the proposed "less-one" strategy, bilingual RNN further improved the translation performance, achieving 0.4 BLEU improvement over RNNLM and 0.8 BLEU improvement over the SMT baseline. Lastly, we combined RNNLM and bilingual RNN trained with the "less-one" strategy and obtained further 0.1 BLEU improvement as shown in Table 2. Currenty, we are training a variety of RNN with right-to-left word order, unsegmented Iraqi Arabic word and so forth for score combination.

Table 3 shows the effect of model initialization using RNNLM against random initialization. Results showed that different initialization led to different translation performance. We suspect that when bilingual training data are insufficient, good model initialization would be crucial since bilingual RNN has more parameters than RNNLM. Insufficient training data may easily lead to local optima.

Table 2. Using monolingual and bilingual RNNLM with 600 hidden nodes on test sets.

Setup	Test1	Test2	Test3	Overall
(1) Baseline	37.0	33.8	34.3	34.7
(2) RNNLM	36.9	33.8	35.3	35.1
(3) bRNN (LSA)	37.4	33.5	35.1	35.0
(4) bRNN (BOW)	37.4	34.1	34.9	35.2
(5) bRNN (BOW, less1)	37.5	34.1	**35.7**	35.5
(2)+(5)	**37.8**	**34.2**	35.6	**35.6**

Table 3. Bilingual RNN (BOW, less1) initialization with 600 hidden nodes on test sets.

Setup	Test1	Test2	Test3	Overall
Random init	37.3	34.0	34.6	35.0
RNNLM init	**37.5**	**34.1**	**35.7**	**35.5**

3.2. Discussion

As a sanity check, we studied the text generation behavior of bilingual RNN on the low-frequent unigrams. Our proposed method can actually boost these rare unigrams statistics when the generated text was constrained by a given source context. For example, an Iraqi word "jyms" (English name James) occurred only 5 times in our training corpus of 4.8 million tokens. If we generate text using bilingual RNN with one source sentence containing an English name "James", we can harvest 308 "jyms" with valid target side context in 5 million generated tokens.

Because we generated segmented Iraq affix streams, we can merge the segmented token back to the original word form to form potentially new valid words. In terms of the new vocabulary from the data generated using bilingual RNN, the size was almost 10 times larger than the monolingual one with same number of generated tokens. Although it was hard for human to check the percentage of all the new words being valid, but we did ask a native Iraqi speaker to check new valid words according to top-100 frequent unigrams in each model; it seems our proposed model generated slightly more new words. The top frequent new words such as "wyjybwn" (and they give) , "nsbh" (ratio) , "EndnA" (we have) with frequency are listed in Table 4. This result shows that bilingual RNN may potentially increase our vocabulary coverage in a more targeted manner than RNNLM due to the text generation constraint using given source sentences. This may facilitate efficient text generation for a target domain when source sentences are available as constraints.

With a close check on the generated text, we do see the quality of the proposed model seemed to over generating the prefixes and hence many invalid words were formed such as "AlEnkm" (roughly like "the on you" in English). Another invalid example is "AlAlAlAlAlAlAlAlAlAlAlAlEly", and it is definitely over generating the prefix "Al". Further study will be needed to impose a grammar to generate meaningful new Iraqi words, to reduce receptions on generating prefixes, and to reduce the likelihood of invalid affix combinations.

Table 4. New valid Iraq words generated by bilingual RNN.

Words	wyjybwn	nsbh	w$AfwA	Al*krt
Freq	120857	76659	66776	44810

We here show one example of translation. Source sentence is: "there will be a delivery time of at least two weeks before we can see the first shipment". Both baseline and RNNLM get the same translations for "delivery" to "wlAdp", which means birth/pregnancy in Iraqi Arabic; while using bilingual RNN gets the correct translation "tslym", which means the shipment, as shown in Table 5. Even though "delivery-wlAdp" was aligned much more stronger than word-pair "delivery-tsylm", bilingual RNN successfully correct this error using the bilingual context.

Table 5. Example of translations.

baseline	rH ykwn Akw *wlAdp* wqt EAlAql AsbwEyn qbl mA nqdr n$wf Awl Hml
RNNLM	rH ykwn Akw *wlAdp* wqt EAlAql AsbwEyn qbl mA nqdr n$wf Awl Hml
bRNN	rH ykwn Akw **tslym** wqt EAlAql AsbwEyn qbl mA nqdr n$wf Awl Hml

4. CONCLUSIONS

In this paper, we investigated bilingual recurrent neural network for reranking N-best lists for statistical machine translation. We compared LSA and bag-of-word representation for a source sentence and found that bag-of-word representation performs better than LSA. We also evaluated our proposed "less-one" strategy to intentionally skip aligned source words at each target word position when constructing a bag-of-word vector. We found that the proposed strategy further improves machine translation performance compared to conventional BOW. Bilingual recurrent neural network has yielded improved re-ranking performance compared to RNNLM across different test sets. In the future, we will integrate more precise source word context beyond the bag-of-word representation and employ GPU to speed up the model training. We also plan to evaluate the proposed approach on different corpora.

5. REFERENCES

[1] M. Sundermeyer, I. Oparin, J. Gauvain, B. Freiberg, R. Schluter, and H. Ney, "Comparison of feedforward and recurrent neural network language models," in *ICASSP*, 2013, pp. 8430–8434.

[2] A. Deoras, T. Mikolov, and K. Church, "A fast rescoring strategy to capture long-distance dependencies," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011, pp. 1116–1127, Association for Computational Linguistics.

[3] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, Montréal, Canada, June 2012, pp. 20–28, Association for Computational Linguistics.

[4] T. Mikolov, K. Martin, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Interspeech*, 2010, pp. 1045–1048.

[5] T. Mikolov, A. Deoras, D. Povey, L. Burget K. Martin, L. Burget, and J. Cernocký, "Strategies for training large scale neural network language models," in *ASRU*, 2011, pp. 196–201.

[6] T. Mikolov, S. Kombrink, L. Burget, J. Cernocký, and S. Khudanpur, "Extensions of recurrent neural network language model," in *ICASSP*, 2011, p. 5528 5531.

[7] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *Spoken Language Technologies*, 2012, pp. 234–239.

[8] T. Mikolov, "Statistical language models based on neural networks," *Ph.D. thesis, Brno University of Technology*, 2012.

[9] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *NAACL*, 2013, pp. 746–751.

[10] H. Schwenk, A. Rousseau, and M. Attik, "Large, pruned or continuous space language models on a gpu for statistical machine translation," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, Montréal, Canada, June 2012, pp. 11–19, Association for Computational Linguistics.

[11] H. Schwenk, "Continuous-space language models for statistical machine translation," in *The Prague Bulletin of Mathematical Linguistics*, 2010, pp. (93): 137–146.

[12] H. S. Le, A. Allauzen, and F. Yvon, "Continuous space translation models with neural networks.," in *HLT-NAACL*, 2012, pp. 39–48.

[13] H. S. Le, T. Lavergne, A. Allauzen, M. Apidianaki, L. Gong, A. Max, A. Sokolov, G. Wisniewski, and F. Yvon, "LIMSI @ wmt12.," in *WMT*, 2012, pp. 330–337.

[14] J. Gao, X. He, W. Yih, and L. Deng, "Learning continuous phrase representations for translation modeling," in *ACL*, 2014.

[15] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," in *ACL*, 2014.

[16] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, October 2013, pp. 1044–1054, Association for Computational Linguistics.

[17] M. I. Jordan, "Graphical models," in *Statistical Science (Special Issue on Bayesian Statistics)*, 2004, 19, pp. 140–155.

[18] K. Yao, G. Zweig, M. Y. Hwang, Y. Y. Shi, and D. Yu, "Recurrent neural networks for language understanding," in *Proceedings of Interspeech*, 2013, pp. 104–108.

[19] Y. C. Tam, I. Lane, and T. Schultz, "Bilingual LSA-based adaptation for statistical machine translation," *Machine Translation*, vol. 21, no. 4, pp. 187–207, December 2007.

[20] D. Chiang, "Hierarchical phrase-based translation," in *Computational Linguistics*, 2007, vol. 33(2).

[21] D. Chiang, K. Knight, and W. Wang, "11,001 new features for statistical machine translation," in *Proc. NAACL–HLT 2009*, 2009, pp. 218–226.

[22] M. Hopkins and J. May, "Tuning as ranking," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011, pp. 1352–1362, Association for Computational Linguistics.

[23] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the ACL-02)*, Philadelphia, PA, July 2002, pp. 311–318.