Improved Neural Machine Translation with POS-Tagging Through Joint Decoding

Xiaocheng Feng[1], Zhangyin Feng[1], Wanlong Zhao[2,3,4(✉)], Nan Zou[5], Bing Qin[1], and Ting Liu[1]

[1] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001, China
[2] Acoustic Science and Technology Laboratory, Harbin Engineering University, Harbin 150001, China
wlzhao@hrbeu.edu.cn
[3] Key Laboratory of Marine Information Acquisition and Security (Harbin Engineering University), Ministry of Industry and Information Technology, Harbin 150001, China
[4] College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin 150001, China
[5] Harbin University of Commerce, Harbin 150001, China

Abstract. In this paper, we improve the performance of neural machine translation (NMT) with shallow syntax (e.g., POS tag) of target language, which has better accuracy and latency than deep syntax such as dependency parsing. We present three NMT decoding models (independent decoder, gates shared decoder and fully shared decoder) to jointly predict target word and POS tag sequences. Experiments on Chinese-English and German-English translation tasks show that the fully shared decoder can acquire the best performance, which increases the BLEU score by 1.4 and 2.25 points respectively compared with the attention-based NMT model.

Keywords: Neural machine translation · Natural language processing · Artificial intelligence

1 Introduction

Neural Machine Translation (NMT) plays an important role in current natural language processing (NLP) community and its performance is usually used as a metric to evaluate the development of artificial intelligence [1]. Recently, deep structure representations (e.g., dependence) are applied to NMT tasks as external features in both encoding and decoding sides, and new architectures have achieved impressive results in translation quality of many language pairs [2–4]. Compared to deep syntax, we favor to shallow structures (e.g., POS tag

S. Han et al. (Eds.): AICON 2019, LNICST 286, pp. 159–166, 2019.
https://doi.org/10.1007/978-3-030-22968-9_14

and chunk) in this work, which have higher accuracy and faster analyzers. We believe that the performance of an NMT system would benefit from POS tag information of target language. Implicit patterns of target language (e.g. word order) could be revealed from the POS tag sequence. For instance, a Chinese POS tagger typically outputs a "*noun*" or "*pronoun*" after an "*adjective*". A desirable English-Chinese translator should follow this protocol to generate Chinese sentences during the translation procedure. Further, a POS tag is more informative than a chunk or a phrase, and is more concise than combinatory category grammar (CCG) supertag [5], which includes about 500 tags.

Following this direction, our work is to examine the benefit of incorporating POS tags on the target-side. Inspired by the success of multi-task learning in NMT [5–7], we develop three encoder-decoder based NMT architectures to improve the performance of NMT, all of which encode the source language sentence into continuous vectors, and then decode the target language sentence and its POS tag sequence. The difference between these model variations is that they gradually share more parameters in decoding process. Concretely, the first model uses two independent decoders to predict both sequences, and the second one shares partial gated units of those two decoders. As for the third one, the decoding layers are fully shared except for two task-specific softmax functions, which are used for generating different target symbols.

We demonstrate the effectiveness of our architectures on Chinese-English and German-English translation datasets. Experimental results show that our proposed models could improve the performance in contrast to single NMT task with the help of target POS tag sequence prediction. Moreover, our best approach (fully shared decoder) outperforms the attention-based NMT model by an average of 1.8 BLEU points on both datasets. Finally, we show that incorporating source-side POS tags into our architectures could achieve improved performance on German-English translation dataset.

2 Standard NMT Model

In this part, we introduce a conventional encoder-decoder architecture for NMT. Generally, the encoder is a recurrent neural network (RNN) with LSTM [8], whose input is a source sentence $\mathtt{x} = [x_1, ..., x_n]$. The decoder is another LSTM-based RNN, which works in a sequential way and generates a word at each time step. The generation of a word is actually selectively replicating a word from the target vocabulary. The probability of generating the word y_t at the t-th time step is calculated as follows, where $f(\cdot)$ is a non-linear function; s_t is the *decoder state* at the time step t, $t \in [1, T]$; $e_{y_{t-1}}$ is the embedding of y_{t-1}; c_t is the context vector, which is calculated by an *attention* model; W is a linear transformation.

$$p_{\mathtt{y}} = \prod_{t}^{T} \mathtt{P}(y_t|\mathtt{x}, y_{<t-1}) = \prod_{t}^{T} \mathtt{softmax}(f(e_{y_{t-1}}, s_t, c_t)W) \tag{1}$$

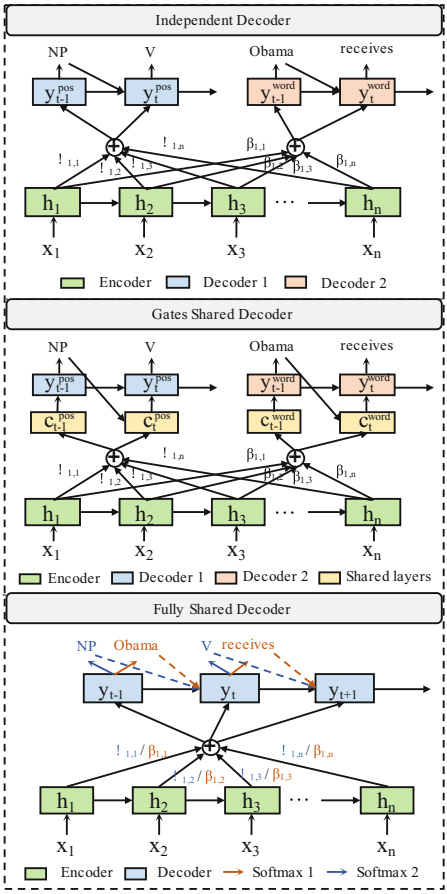

Fig. 1. Architectures of our three NMT models. (Color figure online)

3 Methodology

In this section, we describe the developed neural architectures for NMT. We first introduce a basic multi-task approach, which includes one shared encoder and two different decoders to explicitly model target word and POS tag sequences. Further, we extend two LSTM-based decoders by sharing partial gated neural layers (*input* and *forget*), where the implicit language expression patterns of both sequences are learned. Lastly, we present the third model that shares all decoding layers except for two self-contained softmax functions, where the task-specific and task-related knowledge are modeled together.

3.1 Independent Decoder

An illustration of this model is given in top dashed box of Fig. 1. It shares similar intuition with [6] and [5]. In the multi-task framework [6], the two decoders

are two parameter-independent LSTM-based RNN, which input the same source context representations and output the target language sentence and corresponding POS tag sequence. The two decoders predict a different number of target symbols, resulting in two probability distributions over separate target vocabularies for the words and the POS tags:

$$p_{\mathtt{y}}^{word} = \prod_{t}^{T} \mathtt{P}(y_t^{word} | \mathtt{x}, y_{<t-1}^{word}) \tag{2}$$

$$p_{\mathtt{y}}^{pos} = \prod_{t}^{T} \mathtt{P}(y_t^{pos} | \mathtt{x}, y_{<t-1}^{pos}) \tag{3}$$

3.2 Gates Shared Decoder

The aforementioned multi-task strategy is a loose coupling of the translated words and the POS tags in decoding process. In this subsection, we propose a tighter integration by sharing partial layers of the both decoders. As we know, the LSTM is a special form of recurrent neural networks (RNNs) with three gated units, *input, output* and *forget*, which could control the passing of information along the sequence and thus improve the modeling of long-range dependencies. Actually, we try to share some gated units of the two decoders in order to capture some implicit knowledge between the two tasks for improving the performance of NMT. Among all our six combinations[1], the best performances are achieved when *input* and *output* units are shared and *forget* unit remains independent. An illustration of this model is given in middle dashed box of Fig. 1. The yellow block represents the shared *input* and *output* gates, the blue and pink blocks represent the task-specific units of both decoders, respectively.

3.3 Fully Shared Decoder

In this part, we develop a fully shared encoder-decoder framework. An overview of this architecture is illustrated in bottom dashed box of Fig. 1. The model learns the same set of parameters for modeling the source sentence and predicting the target word and POS tag sequences. Specifically, the decoder needs to be able to predict two sequences of different symbols. Therefore, we equip the shared LSTM-based decoder with two different linear transformation matrices, $W_{word} \in \mathbb{R}^{d \times l}$ and $W_{pos} \in \mathbb{R}^{d \times f}$, where l is the size of target vocabulary and f is the number of target POS tags.

[1] Six combinations (*shared gates / independent gates*): {[*input / forget, output*], [*input, forget / output*], [*input,output / forget*], [*forget / input, output*], [*output / input, forget*], [*forget, output / input*]}.

3.4 Source-Side POS Tags - Shared Embedding

Although our focus is on target-side POS tags, we also experiment with source-side POS tags to show whether the two approaches are complementary. In detail, we follow the previous work [9] and learn a separate embedding for both source-side features such as the word itself and its POS tag. Both feature embeddings are concatenated into one embedding vector which is used in all parts of the encoder model instead of the word embedding [10].

3.5 Training

We used Standford POS tagger [11] to label the corpora instead of using a corpus with gold annotations as in [6]. The final loss was the sum of the losses for the two decoders:

$$loss = -(\log(p_y^{word}) + \log(p_y^{pos})) \tag{4}$$

Models were trained jointly in an alternate manner. We first trained POS tags prediction model and then trained NMT model. The models were optimized using ADADELTA following [9] and all the parameters were initialized randomly with Gaussian distribution. Beam search was adopted for decoding. For brevity, the hyperparameters of the training procedure were given in the published codes[2].

4 Experiments

We conduct experiments on a Chinese-English translation dataset. The training corpora consist of about 1.25 million sentence pairs[3]. We choose the NIST 2002 dataset as our development set, and the NIST 2003, 2004, 2005, 2006 and 2008 datasets as our test sets. Furthermore, we evaluate our model on the IWSLT 2014 translation task of German-English, which consists of sentences-aligned subtitles of TED and TEDx talks. Following the previous study [12], we use 153,000 sentence pairs as training data and extract 6,969 sentence pairs as development set. We also choose dev2010, dev2012, tst2010, tst2011 and tst2012 as test sets, which comprises of about 6,750 sentence pairs.

We compare our models with three strong baselines:

- *RNNSearch*: an in-house implementation of the attention-based NMT system [9] with its default settings.
- *Chunk-Based model*: a chunk-based bi-scale decoder [13] for NMT, in which way, the target sentence is translated hierarchically from chunks to words.
- *CCG Interleaving*: [5] proposed a tight integration in the decoder of the combinatory category grammar (CCG) supertags and the words, where the target sequence includes its CCG supertags as extra tokens.

[2] The codes are implemented with Pytorch, which we plan to release to the community.
[3] The corpora includes LDC2002E18, LDC2003E07, LDC2003E14, the Hansards portion of LDC2004T08, and LDC2005T06.

Table 1. Main experimental results on the NIST Chinese-English and IWSLT German-English translation tasks. * means that the model further incorporates the source-side POS tag information.

Model	Chinese-English							German-English	
	N02	*N03*	*N04*	*N05*	*N06*	*N08*	*Ave.*	*Dev*	*Test*
RNNSearch	36.51	33.32	36.15	33.49	29.77	24.95	31.54	24.93	23.80
Chunk-Based model	36.96	33.83	36.47	33.51	29.82	24.91	31.70	26.57	24.25
CCG interleaving	37.34	34.07	36.95	34.79	30.19	25.85	32.37	27.70	25.50
Independent decoder	37.27	33.90	36.83	33.79	30.25	25.39	32.03	26.83	24.45
*Independent decoder**	37.06	33.25	36.50	33.51	29.93	25.20	31.67	27.61	25.39
Gates shared decoder	37.43	34.27	37.15	**34.99**	30.39	26.05	32.57	26.90	25.04
*Gates shared decoder**	37.12	34.03	36.98	34.08	29.52	25.40	32.01	27.89	26.03
Fully shared decoder	**37.97**	**34.80**	**38.09**	34.73	**30.87**	**26.23**	**32.95**	27.37	25.27
*Fully shared decoder**	37.73	34.03	36.98	34.08	29.52	24.40	32.01	**27.95**	**26.05**

Fig. 2. Case study on the test set.

Table 1 reports the main results of different models measured in terms of BLEU score[4]. Our proposed models outperform different baselines on all sets, which verify that incorporating POS tags in the target-side is helpful for NMT. Our best model *FSD* (*Fully shared decoder*) gains a 1.4 BLEU score improvement upon the standard NMT baseline on Chinese-English corpora and 2.25 (+POS tags in source-side) BLEU points on German-English corpus. Moreover, we find that our proposed three approaches obtain lower scores when incorporating source-side POS tag information on Chinese-English datasets. We speculate that the definition of POS tag in different family of languages may be different. For example, both English and German belong to the Germanic languages while Chinese belongs to the Sino-Tibetan language family.

Furthermore, we also compare *FSD* with the *RNNSearch* and *Chunk-Based model* baselines by subjective evaluation on Chinese-English datasets. Three human evaluators are asked to evaluate the translations of 100 source sentences randomly sampled from the test sets without knowing which system the trans-

[4] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

Table 2. Subjective evaluation results

Model	Adequacy	Fluency
RNNSearch	3.31	3.58
Chunk-Based model	3.43	3.65
Fully shared decoder	3.50	3.84

lation is translated by. The evaluator is asked to give 2 scores: adequacy and fluency, which are from 0 to 5, the larger, the better[5]. Table 2 shows that the subjective evaluation results are highly consistent with the results of objective evaluation. *FSD* improves the two baselines on both the translation adequacy and fluency aspects. Specifically, the fluency increases by an average of 0.225, which confirms the assumption in the introduction that incorporating POS tags in target-side can optimize the word order of the generating sentences. Lastly, we give a case study to illustrate the generated results by *FSD*, as shown in Fig. 2, with a comparison to *RNNSearch* on Chinese-English test set. In the first example, *RNNSearch* generates an extra word "vows", which is addressed in FSD through considering the second POS tag "TO" in predicted POS tag sequence. In the second example, the translated English sentence need to add two definite articles "the". This is partially learned in FSD through taking into account the two "DT" tags in predicted POS tag sequence.

5 Conclusion

In this paper, we focus on NMT task and develop three neural architectures for jointly predicting the target sentence and corresponding POS tag sequence. The basic idea is to guide NMT models towards desired behavior through learning implicit knowledge from target POS tag sequence. Experimental results on Chinese-English and German-English translation tasks have demonstrated that the proposed architectures can significantly improve the translation performance with the help of target POS tag sequence prediction.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
2. Bentivogli, L., Bisazza, A., Cettolo, M., et al.: Neural versus phrase-based machine translation quality: a case study (2016)
3. Eriguchi, A., Tsuruoka, Y., Cho, K.: Learning to parse and translate improves neural machine translation (2017)
4. Hashimoto, K., Tsuruoka, Y.: Neural machine translation with source-side latent graph parsing. In: Proceedings of the Conference on Machine Translation, WMT 2017, pp. 125–135. Association for Computational Linguistics (2017)

[5] The value of kappa is 0.65 in 1–5 scale on two dimensions.

5. Nadejde, M., Reddy, S., Sennrich, R., et al.: Predicting target language CCG Supertags improves neural machine translation. In: Proceedings of the Conference on Machine Translation, WMT 2017, vol. 1, pp. 68–79. Association for Computational Linguistics (2017)
6. Luong, M.-T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. In: Proceedings of the Conference on Machine Translation, WMT 2016. ICLR (2016)
7. Niehues, J., Cho E.: Exploiting linguistic resources for neural machine translation using multi-task learning. In: Proceedings of the Conference on Machine Translation, WMT 2017, vol. 1, pp. 80–89. Association for Computational Linguistics (2017)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: 27th Annual Conference on Neural Information Processing Systems 2013, pp. 3111–3119 (2013)
11. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
12. Ranzato, M., Chopra, S., Auli, M., et al.: Sequence level training with recurrent neural networks. Comput. Sci. (2015)
13. Zhou, H., Tu, Z., Huang, S., et al.: Chunk-based bi-scale decoder for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 580–586. Association for Computational Linguistics (2017)