Learning When to Concentrate or Divert Attention: Self-Adaptive Attention Temperature for Neural Machine Translation

Junyang Lin[1,2], Xu Sun[2], Xuancheng Ren[2], Muyu Li[2], Qi Su[1]
School of Foreign Languages, Peking University[1]
MOE Key Lab of Computational Linguistics, School of EECS, Peking University[2]
{linjunyang, xusun, renxc, limuyu0110, sukia}@pku.edu.cn

Abstract

Most of the Neural Machine Translation (NMT) models are based on the sequence-to-sequence (Seq2Seq) model with an encoder-decoder framework equipped with the attention mechanism. However, the conventional attention mechanism treats the decoding at each time step equally with the same matrix, which is problematic since the softness of the attention for different types of words (e.g. content words and function words) should differ. Therefore, we propose a new model with a mechanism called Self-Adaptive Control of Temperature (SACT) to control the softness of attention by means of an attention temperature. Experimental results on the Chinese-English translation and English-Vietnamese translation demonstrate that our model outperforms the baseline models, and the analysis and the case study show that our model can attend to the most relevant elements in the source-side contexts and generate the translation of high quality.

1 Introduction

In recent years, Neural Machine Translation (NMT) has become the mainstream method of machine translation as it, in a great number of cases, outperforms most models based on Statistical Machine Translation (SMT), let alone the linguistics-based methods. One of the most popular baseline models is the sequence-to-sequence (Seq2Seq) model (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014) with attention mechanism (Bahdanau et al., 2014; Luong et al., 2015). However, the conventional attention mechanism is problematic in real practice. The same weight matrix for attention is applied to all decoder outputs at all time steps, which, however, can cause inaccuracy. Take a typical example from the perspective of linguistics. Words can be categorized into two types, function word, and content word. Function words and content words execute different functions in the construction of a sentence, which is relevant to syntactic structure and semantic meaning respectively. Our motivation is that the attention mechanism for different types of words, especially function word and content word, should be different. When decoding a content word, the attention scores on the source-side contexts should be harder so that the decoding can be more focused on the concrete word that is semantic referent in the source text. But when decoding a function word, the attention scores should be softer so that the decoding can pay attention to its syntactic constituents in the source text that may be several words instead of one word.

To tackle the problem mentioned above, we propose a mechanism called Self-Adaptive Control of Temperature (SACT) to control the softness of attention for the RNN-based Seq2Seq model[1]. We set a temperature parameter, which can be learned by the model based on the attention in the previous decoding time steps as well as the output of the decoder at the current time step. With the temperature parameter, the model is able to automatically tune the degree of softness of the distribution of the attention scores. To be specific, the model can learn a soft distribution of attention which is more uniform for generating function word and a hard distribution which is sparser for generating content words.

Our contributions in this study are in the following: (1). We propose a new model for NMT, which contains a mechanism called Self-Adaptive Control of Temperature (SACT) to control the softness of the attention score distribution. (2). Experimental results demonstrate that our model outperforms the attention-based Seq2Seq model in both Chinese-English and English-Vietnamese translation, with a 2.94 BLEU point and 2.19 BLEU

[1]The code is available at https://github.com/lancopku/SACT

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2985–2990
Brussels, Belgium, October 31 - November 4, 2018.

score advantage respectively[2]. (3). The analysis shows that our model is more capable of translating long texts, compared with the baseline models.

2 Our Model

As is mentioned above, our model is substantially a Seq2Seq framework improved by the SACT mechanism. In this section, we first briefly describe the Seq2Seq model, then introduce the SACT mechanism in detail.

2.1 Seq2Seq Model

We implement the encoder with bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), where the encoder outputs from two directions at each time step are concatenated, and we implement the decoder with unidirectional LSTM. We train our model with the Cross-Entropy Loss, which is equivalent to the maximum likelihood estimation. In the following, we introduce the details of our proposed attention mechanism.

2.2 Self-Adaptive Control of Temperature

In our assumption, due to the various functions of words, decoding at each time step should not use the identical attention mechanism to extract the required information from the source-side contexts. Therefore, we propose our Self-Adaptive Control of Temperature (SACT) to improve the conventional attention mechanism, so that the model can learn to control the scale of the softness of attention for the decoding of different words. In the following, we present the details of our design of the mechanism.

We set a temperature parameter τ to control the softness of the attention at each time step. The temperature parameter τ can be learned by the model itself. In our assumption, the temperature parameter is learned based on the information of the decoding at the current time step as well as the attention in the previous time steps, referring to the information about what has been translated and what is going to be translated. Specifically, it is defined as below:

$$\tau_t = \lambda^{\beta_t} \tag{1}$$

$$\beta_t = \tanh(W_c \tilde{c}_{t-1} + U_s s_t) \tag{2}$$

where s_t is the output of the LSTM decoder as mentioned above, $\tilde{c}_{t-1}$ is the context vector generated by our attention mechanism at the last time step (initialized with the initial state of the decoder for the decoding at the first time step), and λ is a hyper-parameter, which decides the upper bound and the lower bound of the scale for the softness of attention. To be specific, λ should be a number larger than 1[3]. The range of the output value of $\tanh$ function is $(-1, 1)$, so the range of the τ is $(\frac{1}{\lambda}, \lambda)$. Furthermore, the temperature parameter is applied to the conventional attention mechanism.

Different from the conventional attention mechanism, the temperature parameter is applied to the computation of attention score α so that the scale of the softness of attention can be changed. We define the new attention score and context vector as $\tilde{\alpha}$ and $\tilde{c}$, which are computed as:

$$\tilde{c}_t = \sum_{i=1}^{n} \tilde{\alpha}_{t,i} h_i \tag{3}$$

$$\tilde{\alpha}_{t,i} = \frac{exp(\tau_t^{-1} e_{t,i})}{\sum_{j=1}^{n} exp(\tau_t^{-1} e_{t,j})} \tag{4}$$

From the definition above, it can be inferred that when the temperature increases, the distribution of the attention score α is smoother, meaning that softer attention is required, and when the temperature is low, the distribution is sparser, meaning that harder attention is required. Therefore, the model can tune the softness of the attention distribution self-adaptively based on the current output for the decoder and the history of attention, and learns when to attend to only corresponding words and when to attend to more relevant words for further syntactic and semantic information.

3 Experiment

In the following, we introduce the experimental details, including the datasets and the experiment setting.

[2]What should be mentioned is that though the "Transformer" model is recently regarded as the best, the model architecture is not the focus of our study. Furthermore, our proposed mechanism can also be applied to the aforementioned model, which will be a part of our future study.

[3]In our experiments, we use λ of different values, ranging from 2 to 10. The performance differences of models with different λ values are not significant, and we report the results of the model with 4 as the value of λ as it achieves the best performance.

Model	MT-03	MT-04	MT-05	MT-06	Ave.
Moses	32.43	34.14	31.47	30.81	32.21
RNNSearch	33.08	35.32	31.42	31.61	32.86
Coverage	34.49	38.34	34.91	34.25	35.49
MemDec	36.16	39.81	35.91	**35.98**	36.97
Seq2Seq	35.32	37.25	33.52	33.54	34.91
+SACT	**38.16**	**40.48**	**36.81**	35.95	**37.85**

Table 1: **Results of the models on the Chinese-English translation**

3.1 Datasets

Chinese-English Translation We train our model on 1.25M sentence pairs[4] with 27.9M Chinese words and 34.5M English words, and we validate our model on the dataset for the NIST 2002 translation task and test our model on the datasets for the NIST 2003, 2004, 2005, 2006 translation tasks. We use the most frequent 30K words for the Chinese vocabulary and the English vocabulary respectively, covering about 97.4% and 99.7% of the corpora. The evaluation metric is case-insensitive BLEU score computed by `mteval-13a.perl` (Papineni et al., 2002).

English-Vietnamese Translation The training data is from the translated TED talks, containing 133K training sentence pairs provided by the IWSLT 2015 Evaluation Campaign (Cettolo et al., 2015). The validation set is the TED tst2012 with 1553 sentences and the test set is the TED tst2013 with 1268 sentences. The English vocabulary is 17.7K words and the Vietnamese vocabulary is 7K words. The evaluation metric is also BLEU as mentioned above[5].

3.2 Setting

Our model is implemented with PyTorch on an NVIDIA 1080Ti GPU. Both the size of word embedding and the size of the hidden layers in the encoder and decoder are 512. Gradient clipping for the gradients is applied with the largest gradient norm 10 in our experiments. Dropout is used with the dropout rate set to 0.3 for the Chinese-English translation and 0.4 for the English-Vietnamese translation, in accordance with the evaluation on the development set. Batch size is set to 64. We use Adam optimizer (Kingma and Ba, 2014) to train the model[6].

3.3 Baselines

In the following, we introduce our baseline models for the Chinese-English translation and the English-Vietnamese translation respectively.

For the Chinese-English translation, we compare our model with the most recent NMT systems, illustrated in the following. **Moses** is an open source phrase-based translation system with default configurations and a 4-gram language model trained on the training data for the target language; **RNNSearch** is an attention-based Seq2Seq with fine-tuned hyperparameters; **Coverage** is the attention-based Seq2Seq model with a coverage model (Tu et al., 2016); **MemDec** is the attention-based Seq2Seq model with the external memory (Wang et al., 2016).

For the English-Vietnamese translation, the models to be compared are presented below. **RNNSearch** The attention-based Seq2Seq model as mentioned above, and we present the results of (Luong and Manning, 2015); **NPMT** is the Neural Phrase-based Machine Translation model by Huang et al. (2017).

4 Results and Analysis

In the following, we present the experimental results as well as our analysis of temperature and case study.

4.1 Results

We present the performance of the baseline models and our model on the Chinese-English translation in Table 1. As to the recent models on the same task with the same training data, we extract their results from their original articles. Compared with the baseline models, our model with the SACT for the softness of attention achieves

[4]The dataset is extracted from LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06

[5]For comparison with the existing system, we use `multi-bleu.perl` instead.

[6]$\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$

Model	BLEU
RNNSearch	26.10
NPMT	27.69
Seq2Seq	26.93
+SACT	**29.12**

Table 2: **Results of the models on the English-Vietnamese translation**

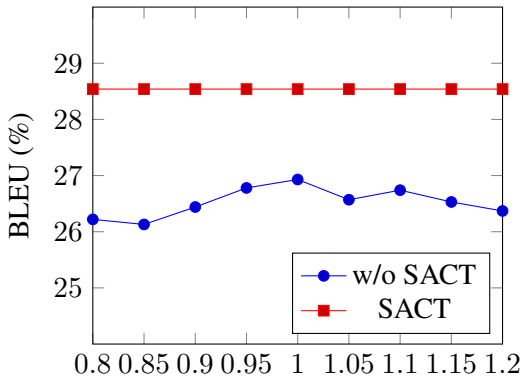


Figure 1: BLEU scores of the Seq2Seq models with fixed values for the temperature parameter. Models are tested on the test set of the English-Vietnamese translation.

better performance, with the advantages of BLEU score 2.94 over the conventional attention-based Seq2Seq model. The SACT effectively learns the temperature to control the softness of attention so that the model can utilize the information from the source-side contexts more efficiently.

We present the results of the models on the English-Vietnamese translation in Table 2. Compared with the attention-based Seq2Seq model, our model with the SACT can outperform it with a clear advantage of 2.17 BLEU score. We also display the most recent model NPMT (Huang et al., 2017) trained and tested on the dataset. Compared with NPMT, our model has an advantage of BLEU score of 1.43. It can be indicated that for low-resource translation, the information from the deconvolution-based decoder is important, which brings significant improvement to the conventional attention-based Seq2Seq model.

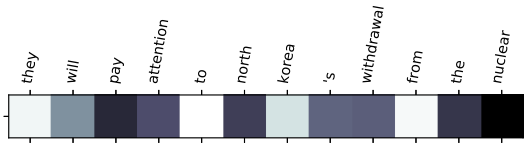
4.2 Analysis

In order to verify whether the automatically changing temperature can positively impact the performance of the model, we implement a series of models with fixed values, ranging from 0.8 to 1.2, for the temperature parameter. From the results shown in Figure1, it can be found that the automatically changing temperature can encourage the model to outperform those with fixed temperature parameter.

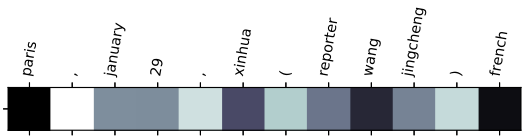
Furthermore, as our model generates a temperature parameter at each time step of decoding, we present the heatmaps of two translations from the testing on the NIST 2003 for the Chinese-English translation on Figure 2. From the heatmaps, it can be found that the model can adapt the temperature parameter to the generation at the current time step. In Figure 2(a), when translating words such as "to" and "from", which are syntactic-relevant prepositions and both lack direct corresponding words in the source text or pronoun such as "they", whose corresponding word "tamen" in the source may be a part of the possessive case or the objective case, the temperature parameter increases to soften the attention distribution so that the model can attend to more relevant elements for accurate extraction of the information from the source-side contexts. On the contrary, when translating content words or phrases such as "pay attention" and "nuclear", where there are direct corresponding words "zhuyi" and "hezi" in the source text, the temperature decreases to harden the attention distribution so that the model can focus on the corresponding information in the source text for accurate translation. In Figure 2(b), the temperature parameters for the punctuations are high as they are highly connected to the syntactic structure and those for the content words with concrete correspondences such as location "paris", name of organization "xinhua", name of person "wang" and nationality "french".

4.3 Case Study

We present two examples of the translation of our model in comparison with the translation of the conventional attention-based Seq2Seq model and the golden translation. In Table 3(a), it can be found that the translation of the conventional Seq2Seq model does not give enough credit to the word "chengzhang" (meaning "growth"), while our model can not only concentrate on the word but also recognize the word as a noun ("chengzhang" in Chinese can be both noun and verb). Even compared with the golden translation, the translation of our model seems better, which is a grammatical and coherent sentence. In Table 3(b), although the Seq2Seq model can generate the



(a)



(b)

Figure 2: **Examples of the heatmaps of temperature parameter** The dark color refers to low temperature, while the light color refers to high temperature.

Source: 中国 大陆 手机 用户 成长 将 减缓
Gold: growth of mobile phone users in mainland china to slow down
Seq2Seq: mainland cell phone users slow down
SACT: the growth of cell phone users in chinese mainland will slow down

(a)

Source: 自 去年 12 以来, 受 委内瑞拉 国内 大罢工 和 伊拉克 战争 的影响, 国际 市场 原油 价格 持续 上涨 。
Gold: since december last year , the price of crude oil on the international market has kept rising due to the general strike in venezuela and the threat of war in iraq .
Seq2Seq: since december last year , the international market has continued to rise in the international market and the threat of the iraqi war has continued to rise .
SACT: since december last year , the international market of crude oil has continued to rise because of the strike in venezuela and the war in iraq .

(b)

Table 3: Two examples of the translation on the NIST 2003 Chinese-English translation task. The difference between Seq2Seq and SACT is shown in color.

translation about the increase in the crude oil, it wrongly connects the increase with the threat of war in Iraq. In contrast, as our model has more capability of analyzing the syntactic structure by softening the attention distribution in the generation of syntax-relevant words, it extracts the causal relationship in the source text and generates the correct translation.

5 Related Work

Most systems for Neural Machine Translation are based on the sequence-to-sequence model (Seq2Seq) (Sutskever et al., 2014), which is an encoder-decoder framework (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014). To improve NMT, a significant mechanism for the Seq2Seq model is the attention mechanism (Bahdanau et al., 2014). Two types of attention are the most common, which are proposed by Bahdanau et al. (2014) and Luong et al. (2015) respectively.

Though the attention mechanism is powerful for the requirements of alignment in NMT, some prominent problems still exist. To tackle the impact of the attention historyTu et al. (2016); Mi et al. (2016); Meng et al. (2016); Wang et al. (2016); Lin et al. (2018a) take the attention history into consideration. An important breakthrough in NMT is that Vaswani et al. (2017) applied the fully-attention-based model to NMT and achieved the state-of-the-art performance. To further evaluate the effect of our attention temperature mechanism, we will implement it to the "Transformer" model in the future. Besides, the studies on the attention mechanism have also contributed to some other tasks (Lin et al., 2018b; Liu et al., 2018)

Beyond the attention mechanism, there are also important methods for the Seq2Seq that contribute to the improvement of NMT. Ma et al. (2018) incorporates the information about the bag-of-words of the target for adapting to multiple translations, and Lin et al. (2018c) takes the target context into consideration.

6 Conclusion and Future Work

In this paper, we propose a novel mechanism for the control over the scope of attention so that the softness of the attention distribution can be changed adaptively. Experimental results demonstrate that the model outperforms the baseline models, and the analysis shows that our temperature parameter can change automatically when decoding diverse words. In the future, we hope to find out more patterns and generalized rules to explain the model's learning of the temperature.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 61673028) and the National Thousand Young Talents Program. Qi Su is the corresponding author of this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. *Proc. of IWSLT, Da Nang, Vietnam*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014*, pages 1724–1734.

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

Po-Sen Huang, Chong Wang, Dengyong Zhou, and Li Deng. 2017. Neural phrase-based machine translation. *CoRR*, abs/1706.05565.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP 2013*, pages 1700–1709.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Junyang Lin, Shuming Ma, Qi Su, and Xu Sun. 2018a. Decoding-history-based adaptive control of attention for neural machine translation. *CoRR*, abs/1802.01812.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018b. Global encoding for abstractive summarization. In *ACL 2018*, pages 163–169.

Junyang Lin, Xu Sun, Xuancheng Ren, Shuming Ma, Jinsong Su, and Qi Su. 2018c. Deconvolution-based global decoding for neural machine translation. In *COLING 2018*, pages 3260–3271.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *AAAI 2018*.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, pages 1412–1421.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018. Bag-of-words as target for neural machine translation. In *ACL 2018*, pages 332–338.

Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive attention for neural machine translation. In *COLING 2016*, pages 2174–2185.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *EMNLP 2016*, pages 955–960.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL, 2002*, pages 311–318.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS, 2014*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL 2016*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*, pages 6000–6010.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Memory-enhanced decoder for neural machine translation. In *EMNLP 2016*, pages 278–286.