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

Presenter: Baosong Yang

Motivation

- The softness of the attention for different types of words should different.

- Function words and content words execute different functions in the construction of a sentence.
  - For content word, attention should be harder--focus on the concrete word that is semantic referent.
  - For function word, attention should be softer--syntactic constituents that may be several words instead of one word.

Self-Adaptive Control of Temperature

The range of τ is $(\frac{1}{\lambda}, \lambda)$

$$\tau_t = \lambda^{\beta_t}$$
$$\beta_t = \tanh(W_c \tilde{c}_{t-1} + U_s s_t)$$

Thus, the scale of the softness of attention can be changed.

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

When the temperature is high, the distribution is smoother

--attend to more relevant words

When the temperature is low, the distribution is sparser

--attend to only coresponding words

Experiments

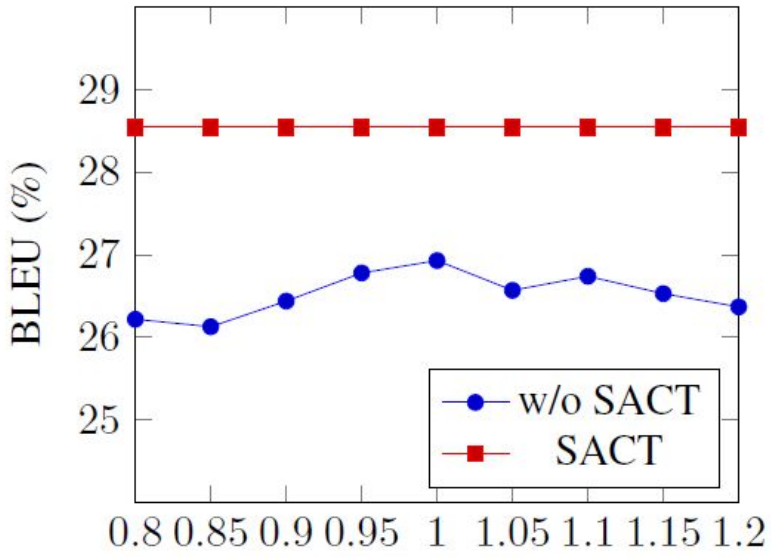
- Zh-En (NIST) 1.25M
- En-Vietnamese (IWSLT) 0.13M

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**

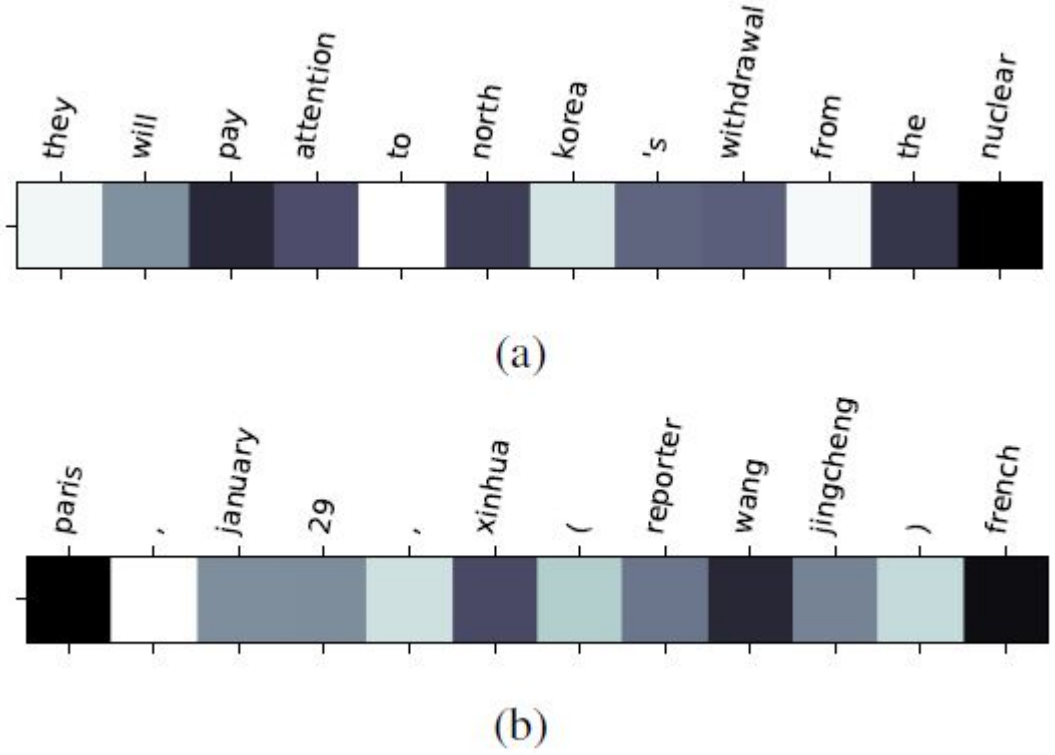
Model	BLEU
RNNSearch	26.10
NPMT	27.69
Seq2Seq	26.93
+SACT	**29.12**

Analysis

Automatically changing temperature can encourage the model to outperform those with fixed temperature parameter.



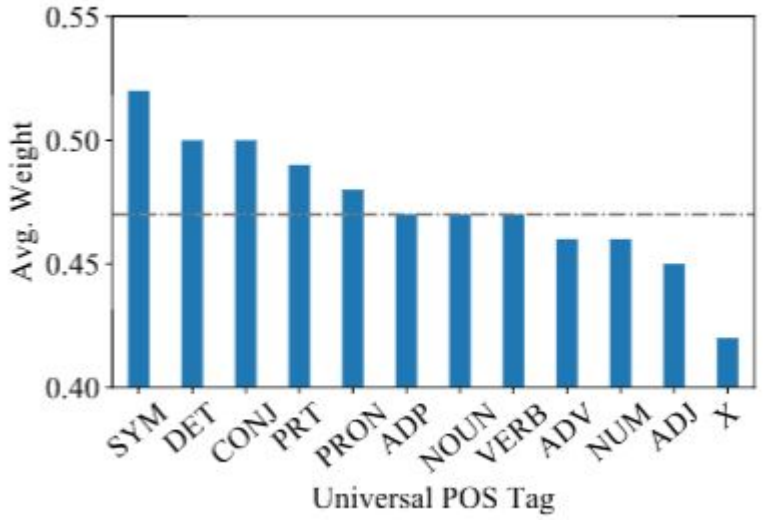
Function words--high temperature
Content words--low temperature



Conclusion

- Interesting motivation, simple but effective approach

- Consistent with our findings on contextural information:
  - Function words require more global information
    - -> mean of representations
    - -> smooth attention distribution
    - -> low temperature (softer attention)



- The concept of temperature is similar to the window size.

- Adopted in self-attention model?

$$\text{ATT}(Q, K) = softmax(energy + G).$$

$$G_{i,j} = -\frac{(j - P_i)^2}{2\sigma_i^2},$$