Word Attention for Sequence to Sequence Text Understanding

[1]Lijun Wu, [2]Fei Tian, [2]Li Zhao, [1]Jianhuang Lai and [2]Tie-Yan Liu

[1] Sun Yat-sen University   [2] Microsoft Research Asia

1. Motivation

- Typical attention mechanism in recurrent neural network (RNNs) builds attention upon **subsequence** representation on source sentence.
- *Word Attention* builds itself upon clean and specific **word-level representation**.
- Enhance the model to extract **more adaptive and comprehensive** source context vectors on different abstractive levels.

2. Contribution

- We leverage source side word level information to form a complementary **attentive word context** besides the hidden context.
- We propose **contextual gates** to dynamically select the amount of hidden context and word context.
- State-of-the-art result on WMT'14 English-French 12M training data

3. Word Attention

- Compute **word attention weights** based on word embedding

$$\beta_{ij} = \frac{\exp(e_{ij}^{\beta})}{\sum_{k=1}^{T_x} \exp(e_{ik}^{\beta})}, \quad e_{ij}^{\beta} = v_b^T \tanh(W_b s_{i-1} + U_b \boldsymbol{x}_j)$$

- Word Context

$$c_i^{\beta} = \sum_{j=1}^{T_x} \beta_{ij} \boldsymbol{x}_j$$

- Update target hidden state and predict next token

$$s_i = f\left(s_{i-1}, y_{i-1}, c_i^{\alpha}, \boldsymbol{c}_i^{\boldsymbol{\beta}}\right)$$

$$p(y_j | y_{<j}, x) = g\left(y_{i-1}; s_i; c_i^{\alpha}; \boldsymbol{c}_i^{\boldsymbol{\beta}}\right)$$

4. Contextual Gates

- **Contextual Gates** to combine hidden and word context

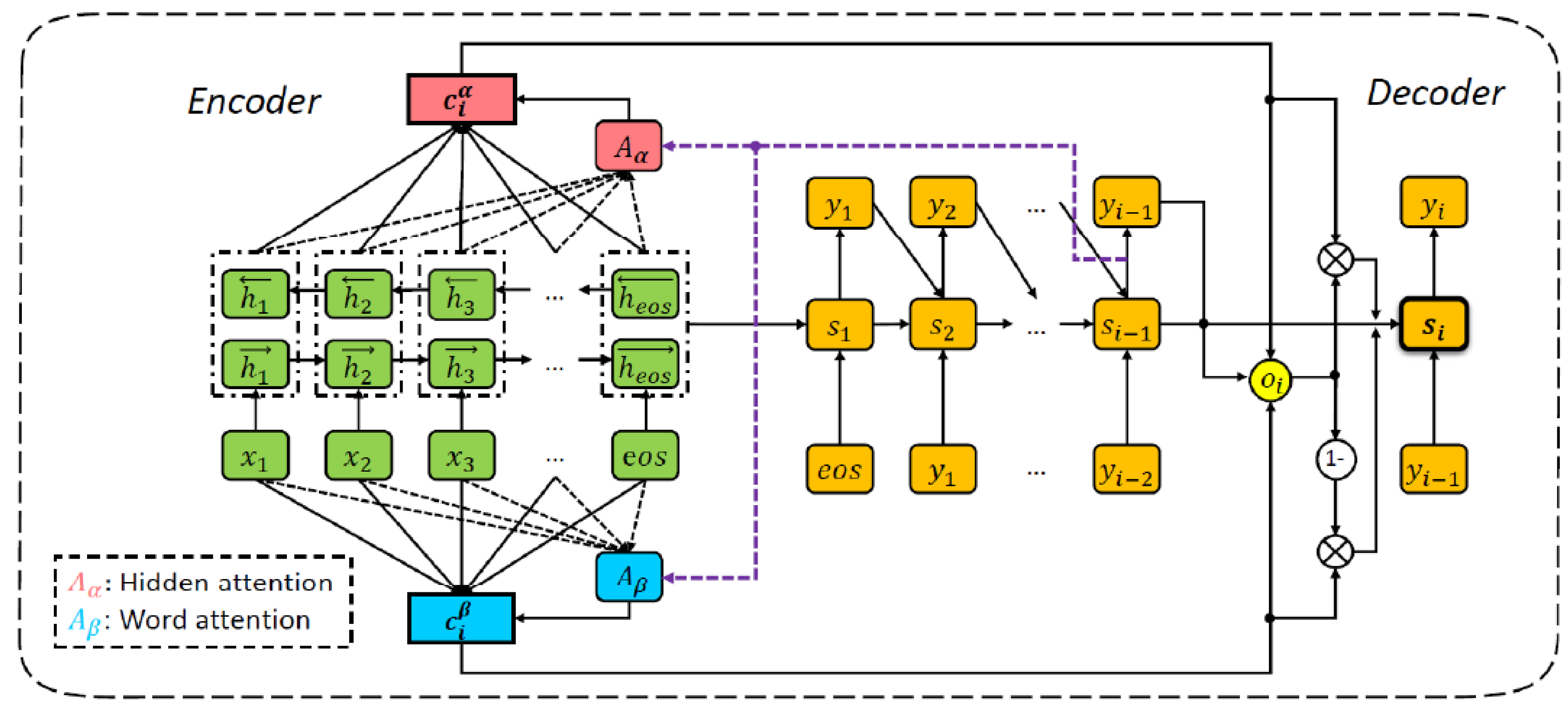
$$\boldsymbol{o}_i = \boldsymbol{\sigma}\left(\boldsymbol{W}_o \boldsymbol{y}_{i-1} + \boldsymbol{U}_o \boldsymbol{s}_{i-1} + \boldsymbol{C}_o^{\alpha} \boldsymbol{c}_i^{\alpha} + \boldsymbol{C}_o^{\beta} \boldsymbol{c}_i^{\beta}\right)$$

$$s_i = f(s_{i-1}, y_{i-1}, \boldsymbol{o}_i \cdot c_i^{\alpha}, (\boldsymbol{1} - \boldsymbol{o}_i) c_i^{\beta})$$

Contact

- wulijun3@mail2.sysu.edu.cn

5. Architecture



6. Experiments

- Text Summarization, Gigaword

Model	RG-1	RG-2	RG-L
ABS	29.55	11.32	26.42
ABS+	29.76	11.88	26.96
RAS-Elman	33.78	15.97	31.15
Feats2s	32.67	15.59	30.64
Luong-NMT	33.10	14.45	30.71
Shen MLE	32.67	15.23	30.56
+MRT	36.54	16.59	33.44
RNNsearch	33.67	15.68	31.67
+Word Attention	35.64	16.64	33.03
+Contextual Gates	**35.93**	**16.99**	**33.41**

Table 1: ROUGE F1 scores on abstractive summarization test set. RG-N stands for N-gram based ROUGE F1 score, RG-L stands for longest common subsequence based ROUGE F1 score. Our work is significantly better than RNNsearch ($p < 0.01$).

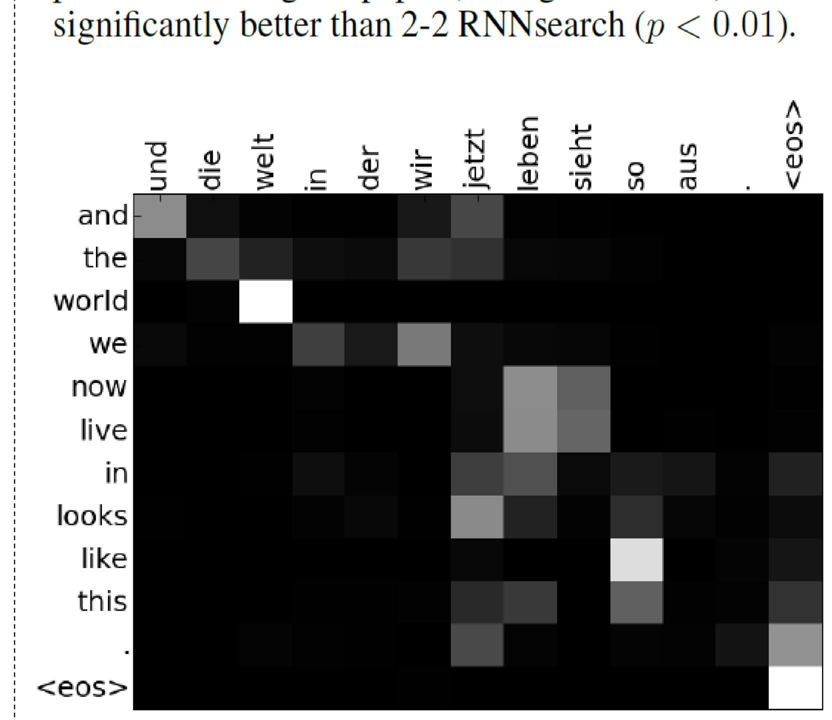
- Neural Machine Translation, WMT'14 En-Fr, IWSLT'14 De-En

Model	Word	Params	BPE	Params
NPMT+LM	29.16	-	-	-
2-2 RNNsearch	29.01	24.3M	31.03	25.0M
+Word Attention	29.68	24.9M	31.71	25.6M
+Contextual Gates	**29.91**	25.6M	**31.90**	26.3M

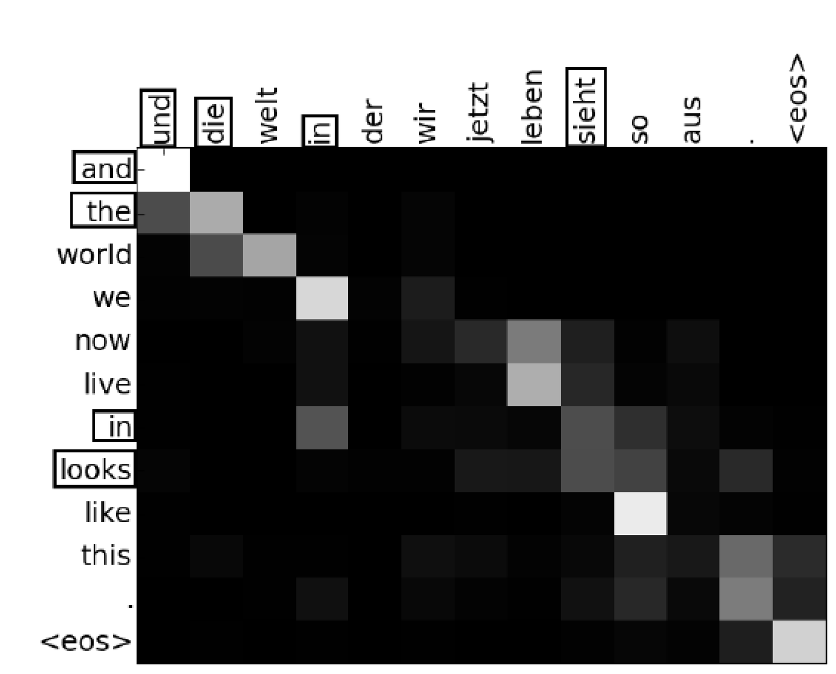
Table 3: BLEU scores on De-En test set for 2-layer models. The BLEU number for baseline model "NPMT+LM" is reported in the original paper (Huang et al. 2017). Our work is significantly better than 2-2 RNNsearch ($p < 0.01$).

Model	Data	BLEU
LAU (Wang et al. 2017)	$12M$	35.10
Deep-Att (Zhou et al. 2016)	$12M$	35.90
Deep-Att (Zhou et al. 2016)	$36M$	37.70
Deep-Att+PosUNK (Zhou et al. 2016)	$36M$	39.20
GNMT (Wu et al. 2016)	$36M$	38.95
4-4 RNNsearch	$12M$	38.50
+Contextual Gates	$12M$	**39.10**

Table 4: BLEU scores on En-Fr test set. Our work is significantly better than 4-4 RNNsearch ($p < 0.05$).



(a) Attention weights from RNNsearch.



(b) Gated attention weights from our model.

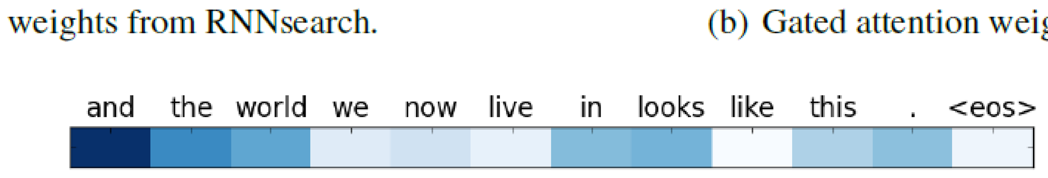


Figure 3: Visualization of the gate units on one De-En translation case. This figure shows the target sentence. The deeper blue color refers to larger value of $1 - o_i$, which means the decoder concentrates more on word context.