Paper Reading: Bag-of-Words as Target for Neural Machine Translation (ACL 2018)

Shuming Ma, Xu Sun, Yizhong Wang, Junyang Lin

Peking University

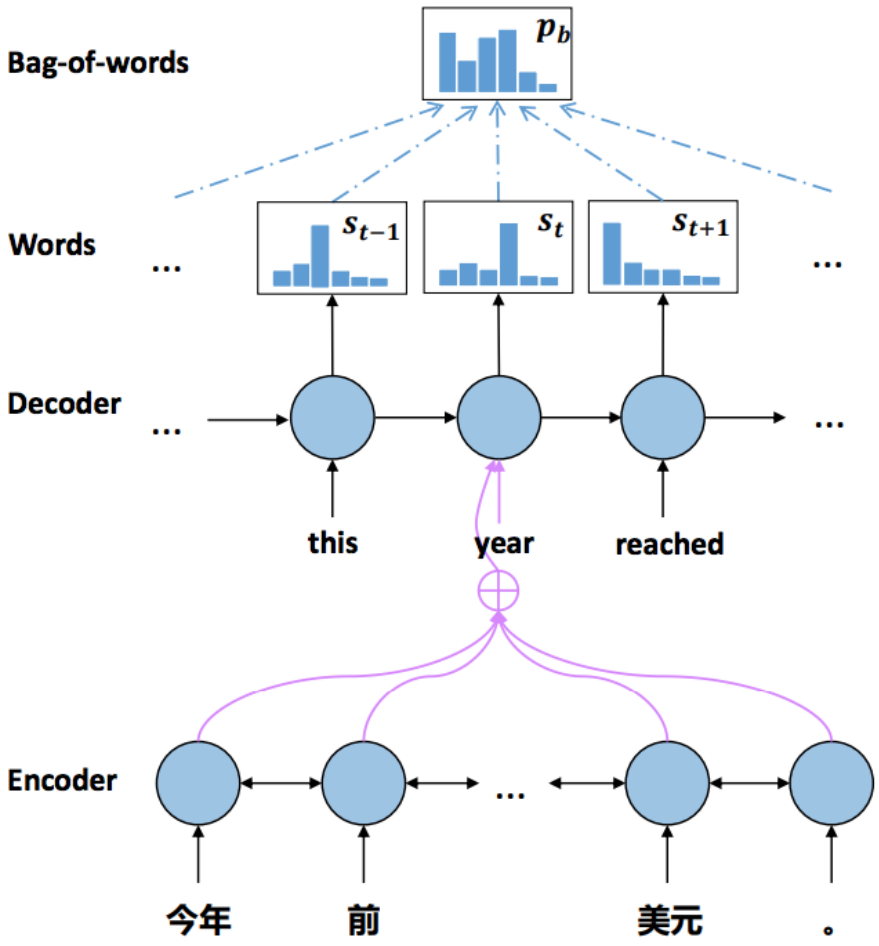
Motivation

- A sentence can be translated into more than one correct sentences, which have different syntax structures and expressions but share the same meaning.
- Current NMT models only have one reference sentences as the targets.
- The generated sentences which cover more words in the bag-of-words are encouraged, while the incorrect sentences are punished.

Source: 今年前两月广东高新技术产品出口37.6亿美元。
Reference: Export of high - tech products in guangdong in first two months this year reached 3.76 billion us dollars .
Translation 1: Guangdong 's export of new high technology products amounts to us $3.76 billion in first two months of this year .
Translation 2: Export of high - tech products has frequently been in the spotlight , making a significant contribution to the growth of foreign trade in guangdong .

Although Translation 1 is much more reasonable, it is punished more severely than Translation 2 by Seq2Seq.

Bag-of-Words Generation



- The word generator is used to compute the probability of each output word at t-th time step:

$$p_{w_t} = softmax(s_t)$$

$$s_t = W_g v_t + b_g$$

Where W_g and b_g are parameters of the generator.

(continue)

- The sentence-level probability of the generated bag-of-words p_b can be written as:

$$p_b = sigmoid(\sum_{t=1}^{M} s_t)$$

where M is the number of words in the target sentence.

Targets and Loss function

Two targets at the training stage: the reference translation (appears in the training set) and the bag-of-words. Two parts of loss functions:

$$l_1 = -\sum_{t=1}^{M} y_t \log p_{w_t}(y_t)$$

$$l_2 = -\sum_{i=1}^{K} b_i \log p_b(b_i)$$

The total loss function is:

$$l = l_1 + \lambda_i l_2$$

Experiments: on the NIST translation task for Chinese-English

Model	MT-02	MT-03	MT-04	MT-05	MT-06	MT-08	All
Moses (Su et al., 2016)	33.19	32.43	34.14	31.47	30.81	23.85	31.04
RNNSearch (Su et al., 2016)	34.68	33.08	35.32	31.42	31.61	23.58	31.76
Lattice (Su et al., 2016)	35.94	34.32	36.50	32.40	32.77	24.84	32.95
CPR (Zhang et al., 2017)	33.84	31.18	33.26	30.67	29.63	22.38	29.72
POSTREG (Zhang et al., 2017)	34.37	31.42	34.18	30.99	29.90	22.87	30.20
PKI (Zhang et al., 2017)	36.10	33.64	36.48	33.08	32.90	24.63	32.51
Bi-Tree-LSTM (Chen et al., 2017)	36.57	35.64	36.63	34.35	30.57	-	-
Mixed RNN (Li et al., 2017)	37.70	34.90	38.60	35.50	35.60	-	-
Seq2Seq+Attn (our implementation)	34.71	33.15	35.26	32.36	32.45	23.96	31.96
+Bag-of-Words (this paper)	**39.77**	**38.91**	**40.02**	**36.82**	**35.93**	**27.61**	**36.51**

Conclusion

- Adding the bag-of-words as the targets of Seq2Seq at the training stage helps a lot.
- Inspiration for current work: add temporal information by using the bag-of-words loss.