Bag-of-Words as Target for Neural Machine Translation

Shuming Ma[1] Xu Sun[1,2] Yizhong Wang[1] Junyang Lin[3]

[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
[3]School of Foreign Languages, Peking University

{shumingma, xusun, yizhong, linjunyang}@pku.edu.cn

Abstract

Task:

Neural Machine Translation

Problem:

A sentence can be translated into more than one correct sentences.

However, most of the existing neural machine translation models only use one of the correct translations as the targets, and the other correct sentences are punished as the incorrect sentences in the training stage.

Proposal:

We propose an approach that uses both the sentences and the bag-of-words as targets in the training stage, in order to encourage the model to generate the potentially correct sentences that are not appeared in the training set.

Example

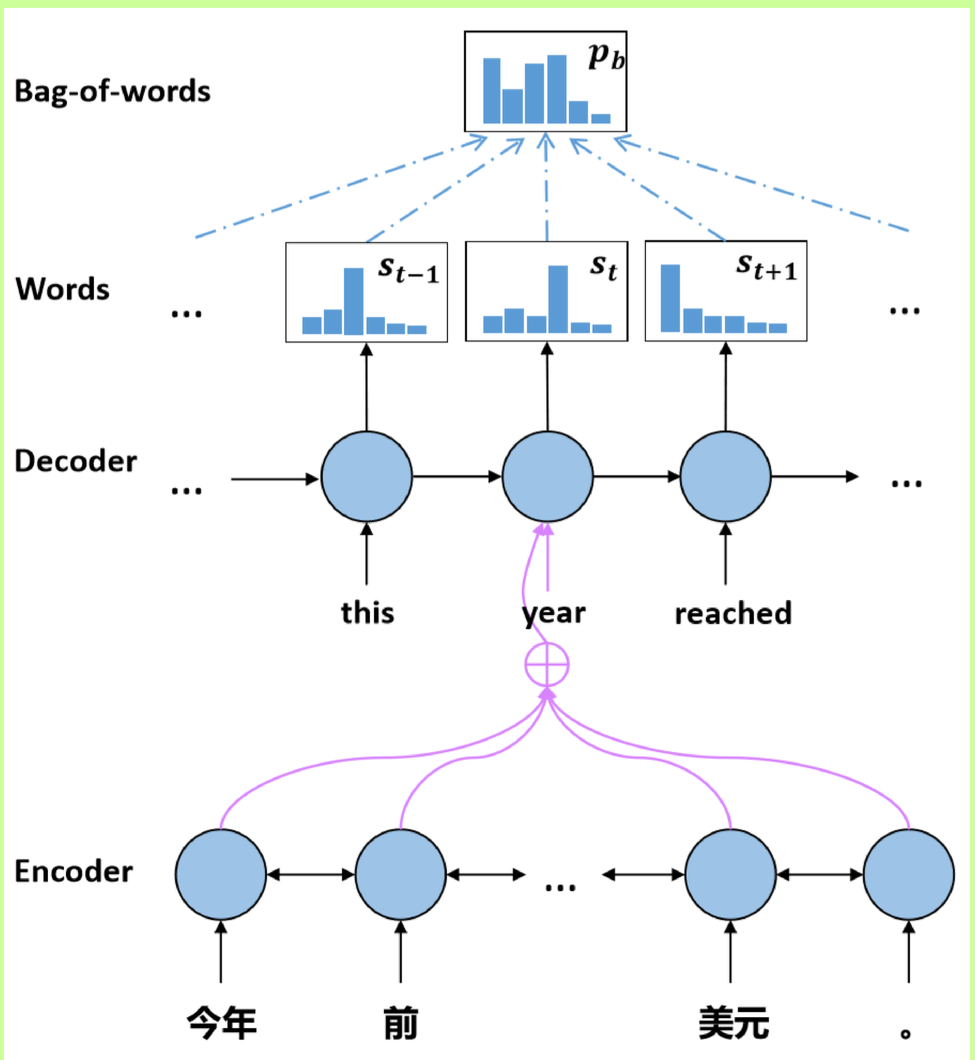
Source: 今年前两月广东高新技术产品出口37.6亿美元。

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 .

Proposed Model



The overview of our model. The encoder inputs the source sentence, and the decoder outputs the word distribution at each position. The distribution of all position is summed up to a sentence-level score, which can be used to generate the bag-of-words.

Targets and Loss Function

The target of sentence:

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

Total loss function:

$$l = l_1 + \lambda_i l_2$$

The target of bag-of-words:

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

Balance two loss functions:

$$\lambda_i = \min(\lambda, k + \alpha i)$$

Datasets

Chinese-English Translation Datset:

Training with 1.25M sentence pairs extracted from LDC corpora

Test on the NIST 2003, 2004, 2005, 2006, 2008 translation tasks

Experimental Results

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**

Results of our model and the baselines (directly reported in the referred articles) on the Chinese-English translation.

Translation Examples

Source: 人类共有二十三对染色体。

Reference: Humans have a total of 23 pairs of chromosomes .

Seq2Seq+Attn: Humans have 23 pairs chromosomes in human chromosome .

+Bag-of-Words: There are 23 pairs of chromosomes in mankind .

Source: 一名奥林匹克筹备委员会官员说:「这项倡议代表筹委会对环保的敏感性。」

Reference: An official from the olympics organization committee said : `` this proposal represents the committee 's sensitivity to environmental protection . ''

Seq2Seq+Attn: An official of the olympic preparatory committee said : `` this proposal represents the <unk> of environmental sensitivity . ''

+Bag-of-Words: An official of the olympic preparatory committee said : `` this proposal represents the sensitivity of the preparatory committee on environmental protection . "

Conclusion

We propose a method that regard both the reference translation (appears in the training set) and the bag-of-words as the targets of Seq2Seq at the training stage. Experimental results show that our model obtains better performance than the strong baseline models on a popular Chinese-English translation dataset. In the future, we will explore how to apply our method to other language pairs, especially the morphologically richer languages than English, and the low-resources languages.