Towards Bidirectional Hierarchical Representations for Attention-Based Neural Machine Translation

Baosong Yang[†], Derek F. Wong[†], Tong Xiao[‡], Lidia S. Chao[†] and Jingbo Zhu[‡]

Presenter: Baosong Yang

[†]NLP^{2}CT Lab, University of Macau
[‡]NiuTrans Lab, Northeastern University

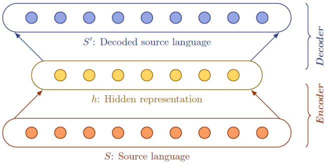
Aug 16, 2017

Contents

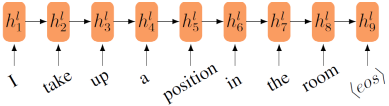
Contents

Neural Machine Translation models

- Encoder-decoder framework
  - Encodes source sentences into a distributed representations.
  - Followed by a decoder generates target translation.

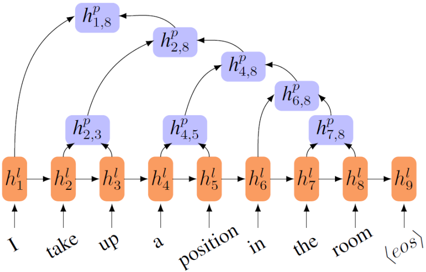


- Traditional sequential encoder
  - Insufficient to fully capture the semantics of a sentence (Tai et al., 2015; Eriguchi et al., 2016).



Conventional Tree-based Encoder

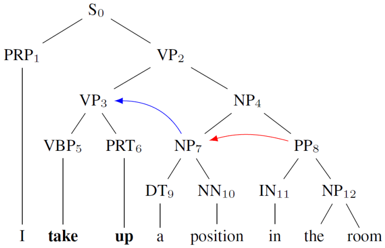
- Tree-based encoder
  - Encodes a source sentence following a syntactic tree (Tai et al., 2015).
  - Tree-to-sequence NMT model is outstanding on structurally distant language pair, e.g. en-jp (Eriguchi et al., 2016).



- Problems:
  - Recursively generate source representations in a bottom-up fashion.
  - The learned representations are limited to local information, while failing to capture the global meaning of a sentence.
  - Ignoring the neighboring context as well as the remote context.

Conventional Tree-based Encoder

- An example:
  - **take up** has the meanings of *start doing something new*, *use space/time*, *accept an offer*, etc.
  - **a position** has the meanings of *location*, *job offer*, *rank/status*, etc.



- The differences in meaning arise as a result of ignoring the neighboring context as well as the remote context, i.e:
  - $h_{\mathrm{NP}_7} \leftarrow h_{\mathrm{PP}_8}$ (sibling)
  - $h_{\mathrm{VP}_3} \leftarrow h_{\mathrm{NP}_7}$ (child of sibling)

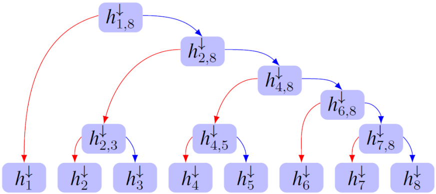
Goal and Contributions

- Goal:
  - Improving tree-based encoder so that the generated source-side representations cover both local and global semantic information.
- Contributions:
  - Bidirectional tree-based encoder
    - To enhance the source-side hierarchical representations
  - Extending to the sub-word level
    - To alleviate the out-of-vocabulary problem
  - A variant weighted tree-based attention mechanism
    - To effectively leverage hierarchical representations

Contents

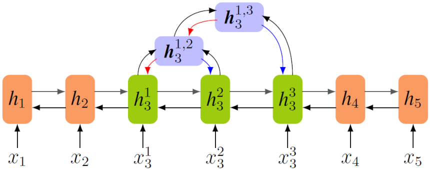
Bidirectional Tree-based encoder

- Bidirectional Leaf-Node Encoding
  - Jointly take into account both preceding and following annotations.
  - $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$
- Bottom-up Encoding
  - Recursively propagated the local context to tree nodes.
  - $h^{\uparrow}_{par} = f_{tree\_gru}(h^{\uparrow}_{l\_child}, h^{\uparrow}_{r\_child})$
- Top-down Encoding
  - Update the representations with global semantic information.
  - $h^{\downarrow}_{l\_child} = f^{l}_{GRU}(h^{\uparrow}_{l\_child}, h^{\downarrow}_{par})$
  - $h^{\downarrow}_{r\_child} = f^{r}_{GRU}(h^{\uparrow}_{r\_child}, h^{\downarrow}_{par})$
  - f^{l}_{GRU} and f^{r}_{GRU} with different parameters are applied to distinguish the left and right structural information.



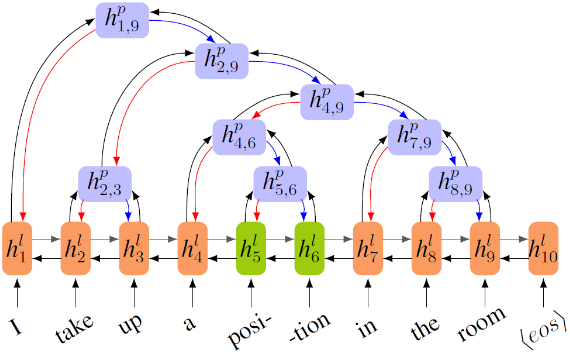
Handling OOV: Tree-based Rare Word Encoding

- Motivation
  - Sequential sub-word encoding: Representing rare words as a sequence of sub-word units (Sennrich et al. 2016).
  - Not applicable to the tree-based NMT model.
- Proposed method



  - 1. Segment the rare word x_i into a sequence of sub-word units $(x_i^1, ..., x_i^n)$ by Byte-pair Encoding (BPE).
  - 2. Built a binary lexical tree by simply composing two nodes in a rightwards fashion, $(((x_i^1, x_i^2), x_i^3)...), x_i^n)$.
  - 3. Sub-word units are encoded following the binary lexical tree.

Bidirectional Hierarchical Encoder



- The vector representations of the sentence, phrases, words as well as sub-word units are therefore based on the global context rather than local information.
- New problem: Attending fairly may cause the problem of over-translation.

Weighted Variant of Attention Mechanism

- To balance the attentive information between the lexical and phrase vectors in the context vector.

$$d_j = (1-\beta_j)\sum_{i=1}^{n}\alpha_j(i)h_i^l + \beta_j\sum_{k=1}^{n-1}\alpha_j(k)h_k^p$$

  - α is the attention score which denotes the correspondence between each source annotation and the current target hidden state
  - $\beta_j \in [0,1]$ manually or automatically weights the expected importance of the representations.
- Gating scalar

$$\beta_j = \sigma(W_\beta c_{j-1} + b_\beta),$$

  - is dominated by the target composite hidden state alone;
  - is a time-dependent scalar;
  - enables the attention model to explicitly quantify how far the leaf and no-leaf states contribute to the word prediction at each time step.

Contents

Experimental settings

- Data set:

Training	Dev	Test		
LDC En-Ch	mt08	mt04	mt05	mt06
1.4M	1,357	1,788	1,082	1,664

- The vocabulary size of the training set:

Training set	Before BPE	After BPE
$\|V\|$ in En	120k	40k
$\|V\|$ in Zh	125k	40k

- NMT setting:

$\|V\|$ size of source	40000
$\|V\|$ size of target	40000
Dimension of word embedding	620
Dimension of leaf layer	512
Dimension of other layer	1024
Batch size	16
Beam search	5
Length of sentence	$<= 40$
Optimizer	AdaDelta

Evaluations on Hierarchical Encoder

- Bidirectional tree-based encoding

Model	BPE	MT04	MT05	MT06	Dev.
tree-based encoder	no	31.90	24.68	24.40	17.63
+ bidirectional leaf-node encoding	no	32.13	24.94	25.02	18.12
+ top-down encoding	no	32.85	25.37	25.30	18.26
hierarchical encoder ($\beta = 0.5$)	no	32.91	25.55	25.52	18.46

- The future context at leaf level can contribute to word prediction.
- The translation quality is improved by considering the global semantic information.

Evaluations on Hierarchical Encoder

- Tree-based rare word encoding

Model	BPE	MT04	MT05	MT06	Dev.
sequential encoder	no	31.26	23.98	24.02	17.20
+ sequential rare word encoding	yes	32.54	25.09	25.07	18.19
+ tree-based rare word encoding	yes	32.56	25.30	24.96	18.33
tree-based encoder	no	31.90	24.68	24.40	17.63
+ tree-based rare word encoding	yes	33.02	25.62	25.24	18.59
hierarchical encoder ($\beta = 0.5$)	no	32.91	25.55	25.52	18.46
hierarchical encoder ($\beta = 0.5$)	yes	33.81	26.47	26.31	19.41

  - Achieves performance comparable to that of the standard BPE in the sequential model, but is applicable to the tree-based NMT model.

Evaluations on Weighted Attention Model

- Four cases:
  - $\beta = \mathbf{0.0}$: Ignore the phrase vectors.
  - $\beta = \mathbf{0.5}$: Non-leaf and leaf vectors participate equally.
  - $\beta = \mathbf{1.0}$: Only consider the phrase vectors.
  - Gating scalar: Dynamically control the proportion.

Model	BLEU	Perplexity	Avg. Length
$\beta = 1.0$	17.16	98.65	21.13
$\beta = 0.5$	19.41	94.73	23.08
$\beta = 0.0$	19.83	94.68	23.33
Gating scalar	**20.10**	**94.18**	23.24

- Phrase representations are unable to fully capture the lexical information of the source sentence. ($\beta = 1.0$)
- Phrase representations tends to generate shorter translation. ($\beta = 1.0$, the average length of Ref. is 23.19.)
- Global information contributes to distinguishing the differences between word meanings (Compare $\beta = 0.5$ with $\beta = 0.0$).
- Through the use of the gating scalar, the hierarchical model achieves progressive improvements.

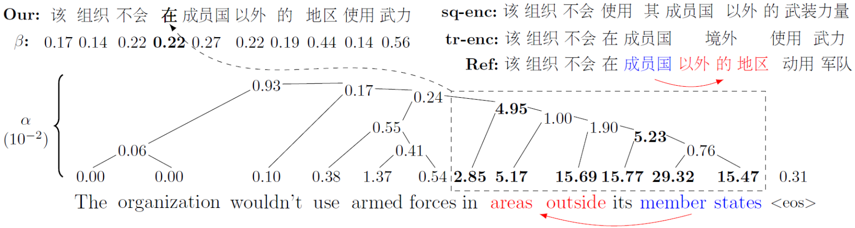
Totally

Model	BPE	MT04	MT05	MT06	Dev.
sequential encoder	no	31.26	23.98	24.02	17.20
+ sequential rare word encoding	yes	32.54	25.09	25.07	18.19
+ tree-based rare word encoding	yes	32.56	25.30	24.96	18.33
tree-based encoder	no	31.90	24.68	24.40	17.63
+ bidirectional leaf-node encoding	no	32.13	24.94	25.02	18.12
+ top-down encoding	no	32.85	25.37	25.30	18.26
+ tree-based rare word encoding	yes	33.02	25.62	25.24	18.59
hierarchical encoder ($\beta = 0.5$)	no	32.91	25.55	25.52	18.46
hierarchical encoder ($\beta = 0.5$)	yes	33.81	26.47	26.31	19.41
+ gating scalar	yes	**34.33**	**26.72**	**26.58**	**20.10**

- Effectively model source-side representations from both the sequential and structural context.
- Outperform conventional models.

Qualitative Analysis

- A translation example.



- Translation examples of sub-words

Source	Reference	Hierarchical	Sequential
liu/jing/min	刘/敬/民 Liú/jìng/mín	刘/敬/民 Liú/jìng/mín	刘/敬/民 Liú/jìng/mín
adventur/er	探险家 Tàn xiǎn jiā	探险家 Tàn xiǎn jiā	探险者 Tàn xiǎn zhě
hi/k/ed	上调 Shàng tiáo	上升 Shàng shēng	发生 Fā shēng

Thank you for listening.