


Improved Neural Machine Translation with Source Syntax

Authors: ShuangzhiWu, Ming Zhou , Dongdong Zhang

Conference: IJCAI 2017

Link: <https://www.ijcai.org/proceedings/2017/0584.pdf>

Abstract

- Researchers have proven that extending word level attention to phrase level attention by incorporating source-side phrase structure can enhance the attention model and achieve promising improvement.
 - However, word dependencies that can be crucial to correctly understand a source sentence are not always in a consecutive fashion (i.e. phrase structure), sometimes they can be in long distance.
 - Phrase structures are not the best way to explicitly model long distance dependencies.
 - This paper proposes a simple but effective method to incorporate source-side long distance dependencies into NMT.
 - Their method based on dependency trees enriches each source state with global dependency structures, which can better capture the inherent syntactic structure of source sentences.
- 

Introduction



- Recently, inspired by the successful application of source-side syntactic information in statistical machine translation (SMT) [Liu et al., 2006], [Eriguchi et al., 2016b] propose a new attentional NMT model which takes advantage of the source-side syntactic information based on the **Head-driven Phrase Structure Grammar** [Sag et al., 1999].
- They align each target word with both source words and source phrases. This kind of extension is effective to handle cases that one target word may correspond to a fragment of consecutive source words.
- However, the long distance syntactic dependencies of the source-side, which can be crucial to correctly understand a sentence, are not explicitly concerned in all previous work.
- Although, in theory, the encoder RNN is able to remember sufficiently long history, we can still observe substantial incorrect translations which are both fluent and grammatical but violate the meaning of source sentences.

- Fig 1 shows incorrect translation example which relates to the source syntax structure.
- Though the translation is grammatically correct, its meaning is inconsistent with the source sentence.
- In Figure 1, if the dependency between the root word “(see the doctor)” and its subject word “(patients)” denoted by a link can be encoded by the NMT encoder, the NMT model is more likely to generate a correct translation.
- This paper tries to solve the above problem by leveraging the source-side dependency tree To explicitly incorporate source word dependencies into NMT framework.
- Based on source dependency trees, each encoder state is enriched from both child-to-head and head-to-child with global knowledge from the dependency structure.



Figure 1: Example of incorrect translation from conventional NMT system. The arrows refer to the dependency link in the dependency tree.

Differences with previous works

The major differences between this work and previous tree based method [Eriguchi et al., 2016b] are in two folds:

- (1) The author models source word relations that are important for understanding source sentences, however, previous work focuses on the mismatch problem that one target word may attend to a source phrase (multiple consecutive words).
- (2) The proposed model enhances the NMT by enriching each encoder state with global source dependency structure, however, previous work improves NMT model by proposing a phrase level attention.

Background



Sutskever et al., 2014 and Bahdanau et al., 2015

- Different from SMT consisting of multiple sub-models, NMT is an end-to-end paradigm directly modeling the conditional translation probability $p(Y | X)$ of the source sentence $X = x_1, x_2, x_3, \dots, x_n$ and the target $Y = y_1, y_2, y_3, \dots, y_m$ with the RNN encoder and the RNN decoder.
- The RNN encoder bidirectionally encodes the source sentence into a separate context vector $H = h_1, h_2, h_3, \dots, h_n$, where $h_i = [\vec{h}_i, \tilde{h}_i]$, \vec{h}_i and \tilde{h}_i are calculated by two RNNs from left-to-right and right-to-left respectively as follows

$$\begin{aligned}\vec{h}_i &= f_{RNN}(x_i, \vec{h}_{i-1}) \\ \tilde{h}_i &= f_{RNN}(x_i, \tilde{h}_{i+1})\end{aligned}$$

where f_{RNN} can be a Gated Recurrent Unit (GRU) [Cho et al.,] or a Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] in practice. In this paper, we use GRU for all RNNs.

Based on target history and the source context, the RNN decoder computes the target translation in sequence by

$$p(Y|X) = \prod_{j=1}^m p(y_j | y_{<j}, H) \quad (1)$$

Typically, for the j th target word, the probability $p(y_j | y_{<j}, H)$ is computed by

$$p(y_j | y_{<j}, H) = g(s_j, y_{j-1}, c_j) \quad (2)$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_j , s_j is the j -th hidden state of decoder RNN, computed by

$$s_j = f_{\text{RNN}}(y_{j-1}, s_{j-1}, c_j)$$

c_j is the source context which is calculated by the attention mechanism. The attention mechanism is proposed to softly align each decoder state with encoder state, where the attention score a_{jk} is computed to explicitly quantify how much each source word contributes to the target word at each time step

$$a_{jk} = \frac{\exp(e_{jk})}{\sum_{d=1}^n \exp(e_{jd})} \quad (3)$$

The calculation for e_{jk} can be in several ways [Luong et al., 2015b], in this paper we compute e_{jk} by

$$e_{jk} = v_a^T \tanh(W_a s_{j-1} + U_a h_k) \quad (4)$$

where v_a , U_a , W_a are the weight matrix

The final source context c_j is the weighted sum of all encoder states

$$c_j = \sum_{k=1}^n a_{jk} h_k \quad (5)$$

The overview of the attention-based NMT is shown in Figure 2.

Although the attention mechanism is effective to model the correspondences between source and target, the long distance syntactic dependencies in the source-side still remain a challenged for a conventional NMT model.

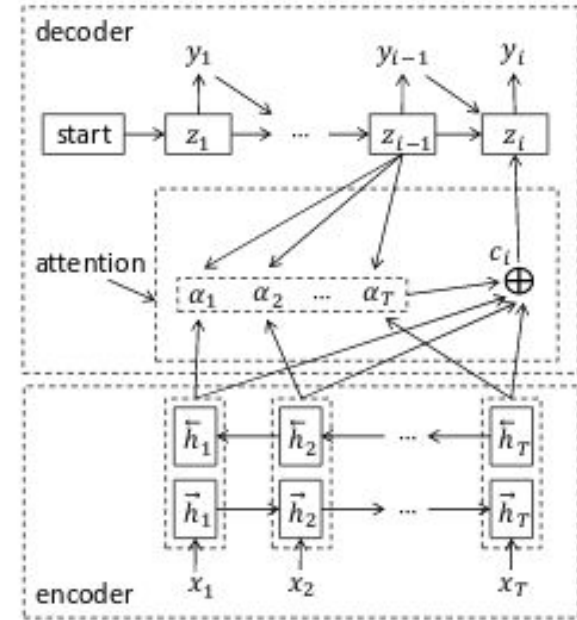


Figure 2: Overview of NMT framework with attention.

The proposed method



Basic Idea

- To incorporate syntactic word relations in NMT, they propose to take advantage of dependency tree.
- Each word in the tree has a parent word which it depends on, except for the root word.
- There are no constituent labels in a dependency tree, the tree directly models word dependencies and syntactic structures of arbitrary distance.
- Given a source sentence $X = x_1, x_2, x_3, \dots, x_n$, where n is the sentence length, and its corresponding dependency tree T , we denote w^h as a possible head node in T , w_l^h as the leftmost child node (or a leftmost subtree) of w^h , w_r^h as the rightmost child node (or a rightmost subtree) of w^h , and w_1^h, \dots, w_j^h as the rest child nodes (or subtrees) of w^h .

All w belongs to X .

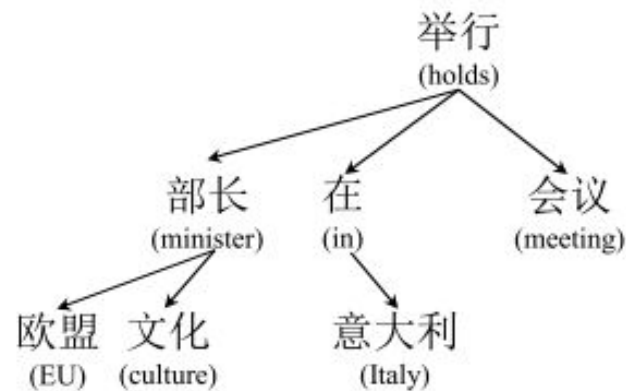
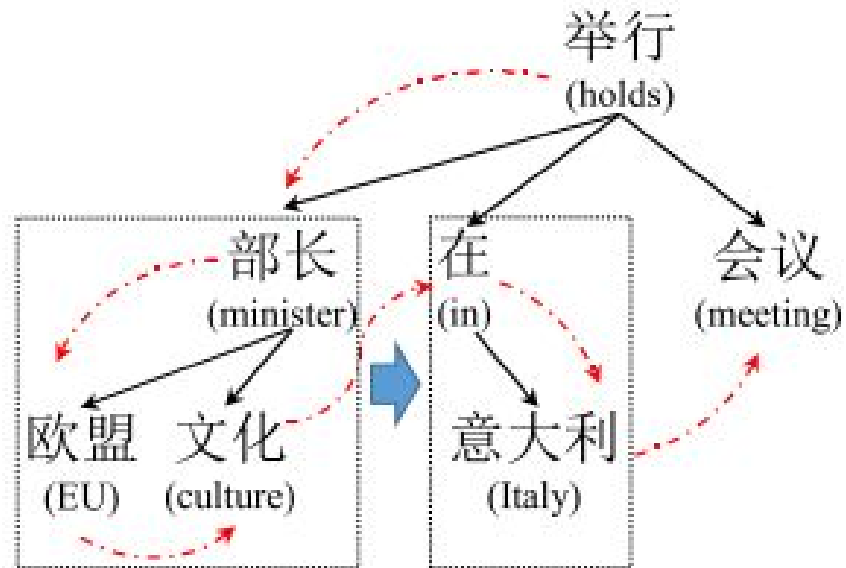


Fig. 3 (a) dependency tree of chinese sentence



Child Enriched Structure (CES)

First way of forming a dependency tree

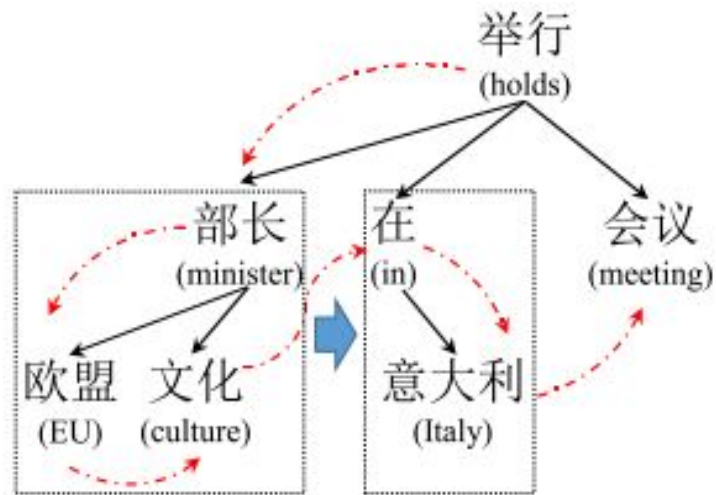
- In most previous work which attempts to leverage syntactic structure in neural networks such as [Tai et al., 2015], a bottom-up fashion is used to construct representations for syntactic trees.
- Good for whole trees, which can facilitate sentiment analysis or similarity of sentences but useless for seq2seq tasks like NMT. The generation of each target word may depend on arbitrary source word.
- Instead of leveraging child nodes to enrich heads, they propose CES to enrich child nodes with global syntactic structures based on the dependency tree.
- Two kinds of context are defined in this structure:

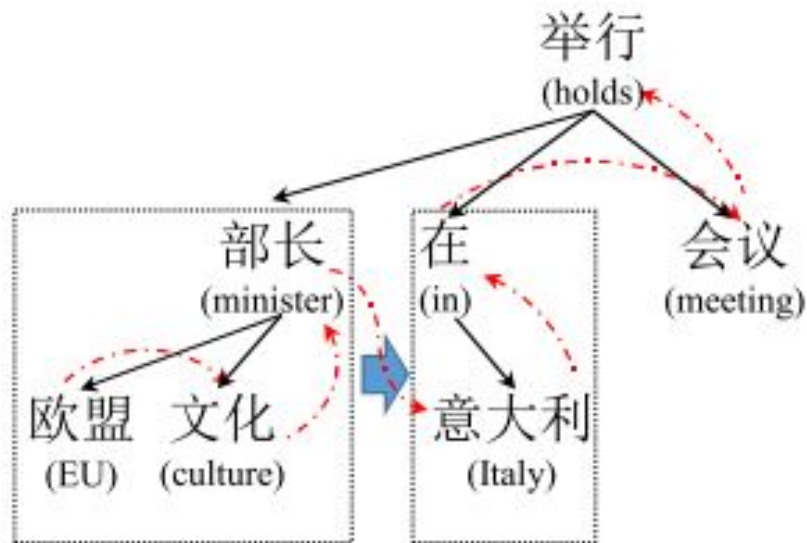
(1) w^h is a direct context for w_l^h .

(2) For a head node w^h , its former child nodes (or subtrees) are contexts for its latter child nodes (or subtrees).

For example w_l^h is a direct context for w_1^h .
 w_1^h is a direct context for w_2^h .

- In the fig, left subtrees are context for right ones.
eg. entire left box should be context for building right box
- To encode this kind of structure in NMT, we use another sequence generated by the pre-order traversal from the dependency tree.





Head Enriched Structure (HES)

Second way of forming a dependency tree

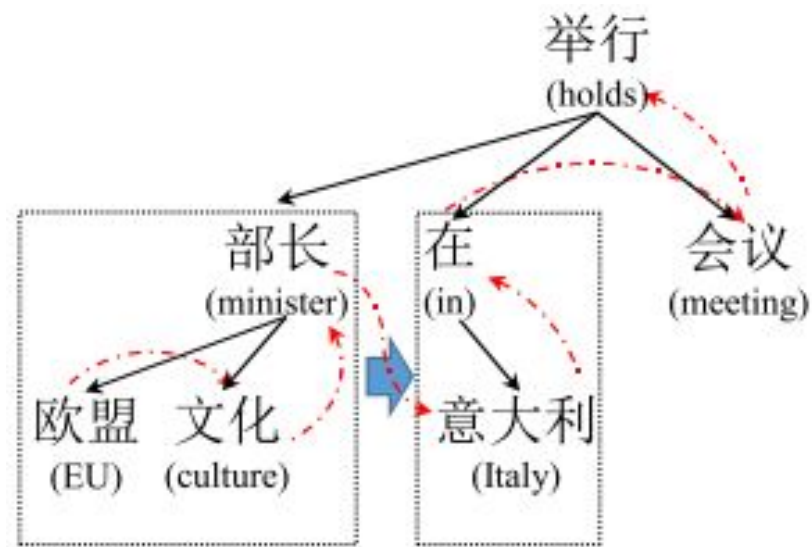
- In addition to child enriched structure, we also enrich the head nodes with its child nodes in the head enriched structure (HES). For this structure, another two kinds of context are defined

(1) The first one is the same with the second one in CES.

(2) w_r^h is a direct context for w^h .

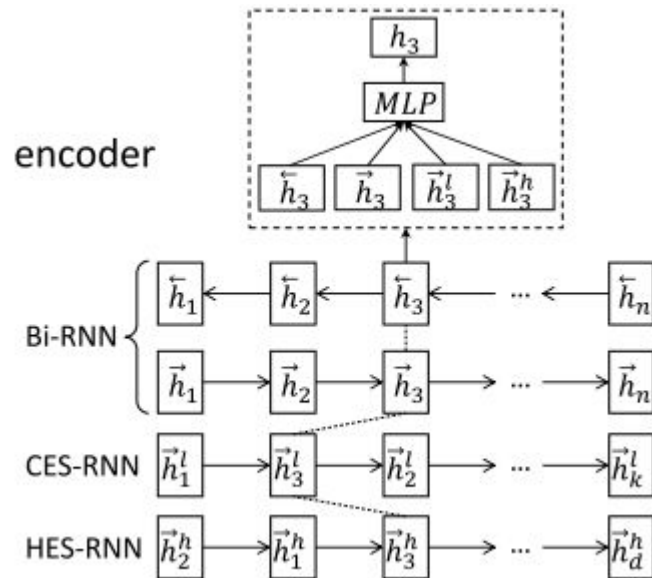
- For the construction of sub-tree in the left box, “(EU)” is first regarded as context for its neighbour “(culture)”, then “(culture)” is used to enrich “minister”.

In addition, the former sub-tree in the left box is context for its neighbor sub-tree in the right box. To encode this kind of structure for NMT, we use another sequence generated by post-order traversal from the dependency tree which perfectly caters to the above description of HES as illustrated by the path of dashed arrows in Figure



The Computation in Encoder

- Two extra RNNs named CES-RNN and HES-RNN are used to encode the two structural sequences in addition (bi-RNN).
- Thus for each source word x_j , we have four hidden state vectors generated by the encoder.
- We denote the two hidden vectors for word x_j from the bidirectional RNNs as \vec{h}_j and \overleftarrow{h}_j , and denote \vec{h}_j^l as the hidden vector from CES-RNN, \vec{h}_j^h as the hidden vector from HES-RNN.



- The final hidden vector h_j used in the decoder is calculated by the four vectors. They are not directly concatenated to avoid the problem of redundancy of information required for decoding.

We apply a MLP function with a smaller hidden size to the four recurrent states before the attention model, as below h_j .

$$h_j = \tanh(W_h \vec{h}_j + U_h \vec{h}_j + V_h \vec{h}_j^l + F_h \vec{h}_j^h) \quad (6)$$

- where W_h , U_h , V_h and F_h are weight matrices. This allows the model to combine the hidden vectors and filter out redundant information.
- The decoder and attention mechanism remains same as the conventional model shown earlier in figure 2.

Conclusion

- This paper proposes simple but effective method to incorporate source dependency structure into NMT encoder.
- Our model can explicitly model word dependencies in the source sentence.
- Experimental results show that our method can achieve promising improvement over the conventional NMT model and outperform the state-of-the-art tree-to-string NMT model.