Paragraph-Parallel based Neural Machine Translation Model with Hierarchical Attention

Yuqi Zhang, Gongshen Liu *
School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, Shanghai, 200240, China
cici--q@sjtu.edu.cn, lgshen@sjtu.edu.cn

Abstract. Neural Machine Translation (NMT) has achieved great developments in recent years, but we still have to face two challenges: establishing a high-quality corpus and exploring optimal parameters of models for long text translation. In this paper, we first attempt to set up a paragraph-parallel corpus based on English and Chinese versions of the novels and then design a hierarchical attention model for it to handle these two challenges. Our encoder and decoder take segmented clauses as input to process the words, clauses, paragraphs at different levels, particularly with a two-layer transformer to capture the context from both source and target languages. The output of the model based on the original transformer is used as another level of abstraction, conditioning on its own previous hidden states. During this process, inter-clause and intra-clause contexts from source and target sides are introduced for translation prediction and both encoder and decoder can profit from contexts in complementary ways. Experimental results show that our hierarchical attention model significantly outperforms six competitive baselines, including ensembles.

1. Introduction
In the past few years, Neural Machine Translation (NMT) has seen great progress, especially in short-single-sentence translation. NMT model is mainly based on the encoder-decoder framework: the encoder compresses the input sentences of the source language into an abstraction from which the decoder generates target sentences. Since in [1], they introduced Multi-Head attention mechanisms to capture contexts in different semantic spaces, the transformer has become a dominant NMT architecture.

However, we still have to face two big challenges. First, the quality of machine translation mainly depends on the quantity and quality of the corpus used. Recently, many corpora that have been well studied are mostly based on TED Talks, Open Subtitles, news and so on. However, it seems that all of the open-source corpora employed in these studies are sentence-aligned even for document-level translation. But in the translation of paragraphs (from novels in our case), though the paragraphs are aligned between source and target languages, there is no strict alignments at sentence level, which renders the models based the above open source corpora less applicable in this situation. Second, the performance of NMT for long text is still not ideal due to the reasons mentioned in [2]. Recently, hierarchical structures used by some researchers [3][16] have shown a clear advantage in modeling paragraphs and documents. In another study, for example, in [11], they segment long sentences into several clauses, and process at word and clause level. Unfortunately, these models cannot be directly used in paragraph parallel corpus since the numbers and orders of sentences in paragraph pairs are different.
To the best of our knowledge, this study is the first attempt to explore end-to-end paragraph NMT based on the paragraph-parallel corpus. Our corpus is established on translated novels, which contributes to addressing the problem of data scarcity in NMT to some extent. To model our corpus, we propose a hierarchical attention model to get context from word-level and clause-level abstractions in a structure and dynamic manner. Specifically, we first segment the input paragraphs into a sequence of clauses by punctuation in source end. Second, we build the coarse parallelism between clauses according to word alignments. Third, we use hierarchical models based on transformers to capture the context from target and source languages. The clause-level abstraction generated by the bottom model will be further used as the input of the top model to generate paragraph-level abstraction. Here, for each predicted word, the context information are modeled by context gate [25] in this hierarchical model, with two attentional mechanisms: one only concentrates on the clause currently being translated (intra-clause context), the other one focuses on all clauses in this paragraph (inter-clause context). In addition, these two contexts can both come from source or target languages. Therefore, the hierarchical encoder can effectively disambiguate the expression of source words, while the hierarchical decoder improves the cohesion and coherence of target words. In this way, our model will jointly optimize the translation of paragraphs, overcoming the difficulties in modeling paragraph-parallel corpus.

Our main contributions are summarized as follows:

- We are the first to introduce paragraph-parallel corpus based on literary works (novels) into NMT. The corpus and code will be shared in GitHub.
- Based on the two layers hierarchical encoder and decoder structure, we capture inter- and intra-clause contexts of target and source languages, which helps to set the proper scope of contexts for prediction.
- According to the experimental results, our model significantly outperforms six strong baselines, in translation tasks.

2. Related Work

In recent years, many researchers have tried to use models with a hierarchical structure in some of different NLP tasks, such as the auto-encoder in paragraph and document [3], translation for long sentences [11], query suggestion [19], dialogue modeling [18], and document classification [23]. Among these, based on our paragraph-parallel corpus, we introduce a two layers hierarchy transformer model, which is beneficial in parameter learning and context modeling in paragraph-translation.

Our work is related to the studies on segmenting long sentences into short ones, and [5] first explored dividing a sentence into a set of parts. Later, many criteria are proposed, such as N-gram, edit distance clues [7], and word alignment [22]. For the translation of long sentences in [11], they adopt a two-level encoder model at word and clause levels. In order to build sentence alignments, in [21], they use a distributed system to reliably mine parallel text from large corpora.

Nowadays, all the existing translation works of long texts are intended to capture contexts from either the source or target side. For document-level translation, in statistical machine translation (SMT), [20] provide a novel method for long distance, sentence-level reordering, and the researchers [10] translate with cross-sentential context. The cache-based approach is also introduced in document-level translation [8]. In NMT, the authors in [16] use a hierarchical attention model to dynamically introduce document-level context into the NMT structure. And the researchers in [15] take both source and target contexts into account using memory networks.

In contrast, to the best of our knowledge, the paragraph-parallel corpus based on literary translations has never been investigated before in NMT. Although, in SMT, the researchers in [21] use e-books as sources for machine translation data sets, while it is based on not end-to-end learning model. The most relevant models could be found include [16] and [11]. But these are always based on sentence aligned corpus, which is not that suitable for the translation of paragraph parallel corpus.
3. The Proposed Approach

Essentially, NMT is to maximize the likelihood of a sentence in the target language as sequence of words \( y = (y_1, y_2, \ldots, y_t) \) when given a sentence in source language in sequence \( x = (x_1, x_2, \ldots, x_n) \), i.e.:

\[
\max \frac{1}{N} \sum_{n=1}^{N} \log (P_\theta(y^n|x^n)) \quad (1)
\]

Thus, paragraph PARA translation is the combination of translating each individual clauses. Specially in this paper, we take into consideration the co-relations among all the clauses within a paragraph from both source and target languages.

\[
\max \frac{1}{N} \sum_{n=1}^{N} \log (P_\theta(y^n|x^n, PARA_y^x, PARA_x^y)) \quad (2)
\]

where \( PARA_y^x = (x^1, x^2, \ldots, x^n) \) and \( PARA_x^y = (y^1, y^2, \ldots, y^T) \) represent the clauses from source and target sides respectively. The contexts of \( PARA_y^x \) and \( PARA_x^y \) are constructed by the hierarchical encoder and decoder.

3.1. Paragraph Segmentation Model

In order to establish the corpus, we need to convert the formats of the bilingual e-books from pdf, mobi, epub, and azw3 to text, remove invalid words and scrambles, divide the bilingual texts into two separate single-language files, and finally manually check and rearrange the sentence order within paragraphs to form the one-to-one correspondences between paragraph pairs.

Based on the proposed corpus, we need to build an approximate correspondence between bilingual sentence pairs in the paragraphs to fit the word-clause-paragraph structure. For this purpose, we run Giza++ [2] to get word alignments, and divide paragraphs into bilingual blocks according to the punctuation (such as comma, period, question mark, slight pause mark, etc.) in the source language. Finally, neighboring blocks with cross-word alignments are combined into a larger block for they cannot be translated separately. For example, as shown in figure 1, the parallel paragraphs are respectively divided into 5 parts each, though the one in the target language can be divided into 6 parts. And the order of these blocks is rearranged to match with the word alignments. It needs to be noted that we choose to divide the paragraphs into clauses rather than natural sentences, because clauses are usually single semantic blocks that could be easily matched between the original text and the translation, whereas natural sentences are more semantically complex and such correspondence is more difficult to establish between texts.

Figure 1. An example of parallel paragraph with word alignments. The dotted line frames represent the boundary of segmented clauses. Index of bilingual blocks from source to target language: 1→5, 2→1, 3→3, 4→4, 5→2.
3.2. Hierarchical Transformer Network

The input paragraph $\text{PARA}$ is divided into $T$ clauses, $\text{PARA} = (c^1, c^2, ..., c^T)$, and $c^j$ is made up of a certain number of words. Specifically, $<\text{eoc}>$ is appended at the end of each clause. As shown at the bottom of figure 2, we use the hierarchical encoder to model the input paragraph, and the structure of the decoder is similar to it.

We build our model based on the transformer model [1] for its high efficiency and accuracy in translation tasks. First, the bottom layer of our model operates at the word level, and generates abstraction of each clause $j$ into a vector $c_j$.

$$Q_w = F_w(h_i),$$

$$C^i = \text{MultiHead}(Q_w, h^i_t),$$

where $h_i$ is the last hidden state of the word to be encoded or decoded at time $t$, $h^i_t$ is the last hidden state of word $i$ of clause $j$. Function $F_w$ is a linear transformation to get query $Q_w$. The Multi-Head attention function [1] can obtain different semantic information within the clause. The hidden representations $h^i_t$ is used as value $V$ and key $K$ for this attention.

And then the top layer of model takes these clauses abstraction as input, and works at the clause level to obtain the abstraction of the entire input paragraph as $\text{PARA}$, at time $t$.

$$Q_c = F_c(h_i),$$

$$\text{PARA}_t = \text{FFN}(\text{MultiHead}(Q_c, c^j)),$$

where $F_c$ is a linear transformation, $Q_c$ is the query, $\text{FFN}$ is a position-wise feedforward layer [1]. Each layer is followed by a normalization layer in transformer.
Different from normal encoder or decoder, our model translates segmented clause with two kinds of attention: an intra-clause context $h_j$ that only focuses on the clause currently being translated, and an inter-clause context $\text{PARA}_i$ that pays attention to all source clauses of the input paragraph. During the encoding and decoding, when the `<eoc>` is detected at the end of a clause, the translation process of this clause terminates, and it moves on to the next clause with a new $h_{t+1}$.

### 3.3. Context Gating

When translating clause $j$, we choose to use a gate [25] to model the context $h_i$ at clause-level and $\text{PARA}_i$ at paragraph-level. Different words have different scopes of context for translation:

$$
\lambda_t = \sigma(W_h h_t + W_{\text{PARA}} \text{PARA}_i), \quad (7)
$$

$$
\tilde{h}_t = \lambda_t h_t + (1 - \lambda_t) \text{PARA}_i, \quad (8)
$$

where $W_h$, $W_{\text{PARA}}$ are parameter matrices, and $h_i$ is the final hidden representation for word $x_t$ or $y_t$ (in source or target language).

### 3.4. Synthesis Model

When encoding or decoding a word, we can take the contexts from different scopes. The contexts are distinguished by the input query $Q$ and value $V$ of the function. In this study, five kinds of context are experimented: one in encoding, three in decoding, and one combining both. In the process of encoding, query is the function of the hidden state $h_x$ of the word $x_t$ currently being encoded in the source side, and values are the states of all the encoded clauses in the same paragraph $h^{\text{H-TRANS encoder}}_y$. In the process of decoding, query is the function of $h_y$ of the currently decoded word $y_t$ in target side, while the values can be of three states: the encoded states of $h^{\text{H-TRANS decoder source}}_x$; the decoded states of clauses in target language $h^{\text{H-TRANS decoder}}_y$; the alignment vectors $b^i$ (H-TRANS decoder alignment). Finally, our model (referred to as H-TRANS-joint) with the combination of hierarchical encoder and decoder is used to capture contexts from both target and source sides.

Notably, the non-segmented paragraphs can also be better translated with only one hidden state in the top transformer layer, as the overall parameters of our model have been optimized by training segmented paragraphs (as proved in Section 4.4). Thus, our model is suitable for all the paragraph pairs.

### 4. Experiments

#### 4.1. Corpus

| Language Pairs (English-Chinese) | Training | Validation | Test |
|---------------------------------|----------|------------|------|
| Number of parallel paragraph pairs | 45.6K | 5.7K | 5.7K |
| Number of segmented clause pairs | 186.1K | 26.5K | 24.3K |
| Kept parallel pairs ≤250 % | 99.5 % | 99.1 % | 98.7 % |
| Segmented paragraphs | 67.5% | 74.6% | 66.2% |
| Average number of clauses/paragraph | 4.08 | 4.64 | 4.27 |
| No crossed word alignments in clauses % | 80.1% | 77.5% | 82.3% |

We establish our paragraph-parallel corpus based on more than one hundred translated novels from English to Chinese, such as *The Wonderful Wizard of Oz*, *Robinson Crusoe*, *Little Women* etc., with a total of 57k bilingual paragraph pairs of 9.7M Chinese words and 10.1M English words. The building of the corpus involves a series of seemingly trivial tasks including converting the formats of the e-books, separating English and Chinese texts from the original bilingual contents, and swiping out the garbles generated during the process. The most troublesome is to manually rearrange the paragraph-pairs of the original texts and their translations as they are not strictly aligned at paragraph level. With our great effort, the paragraph-parallel corpus based on more than one hundred novels are built and it will be shared in GitHub.
In order to better evaluate our model, we use MT track from TED Talks of IWSLT 2017 [6] which contains transcripts of TED talks aligned at sentence level. Each talk is considered to be a paragraph here. We take tst2016-2017 (En-Zh) for testing and the rest for development.

4.2. Setup
We randomly set the ratio of training, verification, and test to 8:1:1 as shown in table 1. We use the case insensitive 4-gram BLEU score [17] to evaluate the results and the script from Moses [14] to test the BLEU scores. The vocabulary size of Chinese and English is 50,000, and the words outside the vocabulary are marked as “unk”. In addition, we keep the paragraph pairs with less than 250 words, covering 99.5% of our corpus, where the max clauses-number of paragraphs is 10. And 81.8% of the input paragraph has less than 80 words of each. We use GIZA++ [2] to align the words with the “grow-diag-final-and” option. And then segmenting paragraph of the training set according to punctuation and word alignments. The sentences for the validation and test sets are segmented based on punctuation. From table 1, we can find most punctuation 80.1% can be used “safely” when dividing source paragraphs into clauses with no crossed word alignments. In addition, from [11], we found the segmentation based on punctuation has a better performance than the maximum entropy classifier. Thus, in this paper, we prefer to use punctuation to do the segmentation.

We used the Open NMT [1] implementation of the transformer. The encoder and decoder are each made up of 6 hidden layers. All hidden states have a dimension of 512, dropout of 0.1 and heads of 8 for Multi-Head attention. The optimization and regularization methods were the same as proposed by [1]. We trained the models in two steps: first, the network parameters are optimized without considering the inter-clause attention, and then the parameters of the whole network are optimized.

Table 2. BLEU score for the different model based on our paragraph parallel English-Chinese corpus and English-Chinese TED Talks.

| Models                        | Our corpus          | TED Talks         |
|-------------------------------|---------------------|-------------------|
| TRANS (only intra)            | 24.10 (baseline)    | 17.07 (baseline)  |
| TRANS (no segmentation)       | 17.12 (-6.98)       | 16.87 (-0.20)     |
| Bi-LSTM (attention)           | 19.62 (-4.48)       | 16.01 (-1.06)     |
| Deep Conv+LSTM [24]           | 21.03 (-3.07)       | 16.13 (-0.89)     |
| H-RNN Search [11]             | 24.52 (+0.42)       | 17.36 (+0.34)     |
| HAN encoder-decoder [16]      | 25.27 (+1.17)       | 17.79 (+0.92)     |
| H-TRANS-joint                 | 26.22 (+2.12)       | 18.41 (+1.34)     |

4.3. Overall Performance
Our model referred to as H-TRANS-joint, is based on H-TRANS encoder and H-TRANS decoder to translate sentences considering the context information from source and target sides. It is compared to the following systems.

1. TRANS (only intra): based on normal transformer to translates the segmented clause independently only with intra-clause contexts.
2. TRANS (no segmentation): based on normal transformer with the whole paragraphs as inputs to train the model.
3. Bi-LSTM (attention): base on bidirectional LSTM model with input as segmented clauses.
4. Deep Conv+LSTM [24] is made up of deep Convolutional Network as encoder and LSTM decoder with input as segmented clauses.
5. H-RNNSearch [11]: based on two layers GRU as encoder to translate segmented clauses sequentially.
6. HAN encoder-decoder (document-level NMT) [16]: based on two layers transformer as encoder and decoder, with the context of previous three clauses.

The overall experimental results of different models are evaluated by the BLEU score, based on our paragraph-parallel corpus and the document-level corpus from TED Talks. As shown in table 2, H-TRANS-joint significantly outperforms TRANS (intra), TRANS (no segmentation), Bi-LSTM (attention),


Deep Conv+LSTM, H-RNN Search by 2.12 and HAN encoder-decoder, by 9.1, 6.6, 5.2, 1.7 and 0.95 BLEU on our corpus, respectively. In addition, our models also have better performance over other six baselines based on TED talks.

Noticeably, compared with the one without segmentation, the paragraph segmentation in our corpus plays an important role in enhancing the translation performance. Furthermore, based on the same segmented clauses, the transformer performs better than bidirectional LSTM and deep convolutional model.

Next, the test sets are divided into different groups according to the clauses number as shown in figure3. We found that our model always outperforms other baselines when the input contains more than three clauses. And the performance of HAN encoder-decoder remains stable. Particularly, compared with TRANS (intra), our model gains 2.25 BLEU points with no less than 10 clauses.

These results strongly prove that our model can better handle paragraph translation, while remains effective in dealing with short sentences and non-segmented paragraphs. Because many long paragraphs have been divided into clauses, the parameters of model are fine-tuned by these training data.

![Figure 3. BLEU scores on different translation groups divided by the clause-number in source language based on our corpus.](image)

4.4. Analysis on the Effect of Attention Mechanism
As shown in table 3, we try to compare the influence of different contexts as mentioned in section 3.4. Obviously, the model with joint hierarchical encoder and decoder gets the best scores, which is significantly higher than the others. An important improvement comes from H-TRANS-encoder because the source language always contains the correct information, while the target language may have incorrect predictions. Using inter-clause context in decoder also improves the performance in translation. In addition, combining H-TRANS-encoder and H-TRANS-decoder can further improve the translation performance, which proves that we can get more information in a complementary way. The abovementioned three kinds of contexts in decoder perform in a similar way.
Table 3. BLEU scores for the different attention mechanisms.

| Models                  | Our corpus | TED Talks |
|-------------------------|------------|-----------|
| TRANS (only intra)      | 24.10 (baseline) | 17.07 (baseline) |
| H-TRANS - encoder      | 25.38 (+ 1.28) | 18.12 (+ 1.05) |
| H-TRANS - decoder      | 21.03 (+ 1.19) | 17.93 (+ 0.86) |
| H-TRANS - decoder (source) | 25.22 (+ 1.12) | 18.06 (+ 0.99) |
| H-TRANS - decoder (alignment) | 25.16 (+ 1.06) | 17.87 (+ 0.80) |
| H-TRANS - joint        | 26.22 (+ 2.12) | 18.41 (+ 1.34) |

4.5. Case Study

Table 4. A translation example based on our corpus. Texts in red are incorrectly translated, in yellow are incoherently translated, in blue are correctly translated.

**Source**

And her insinuation that I was somehow abnormal because I hadn't yet been kissed infuriated me. None of my friends had boyfriends yet. The only girl at the Seattle Academy of Academic Excellence with any dating experience was Wendy Stupacker, who discovered boys in sixth grade—which certainly hadn’t helped her procrastination any. Photographic memory and photogenic looks—tough life.

**Reference**

还有她含沙射影的说我不好,就是因为没接过吻。气死我了。我的朋友们也没男朋友呢。整个西雅图英才预科学校里,唯独一个女孩子有过约会经历——温迪司徒派克。她从六年级就开始接触男孩子了—竟然没影响成绩。她有过目不忘的记忆，和一张让人过目不忘的面—老天不公啊。

**TRANS (only intra)**

她暗示我不是正常的,因为我没有亲过。让我很生气。我的朋友没有一个有男朋友呢。在学院中唯一的女孩在西雅图的任何的约会是温迪 UNK，她发现男孩在六年级,肯定没有帮她把摄影的东西。镜像镜头的外观和艰难的生活。

**TRANS (no segmentation)**

而她批评我很不寻常，因为我从来没有吻过 UNK，我的朋友们都没有幽灵。只有浪漫的女孩在西雅图的 UNK UNK 大学经历学习成绩优秀睁开眼睛，他 UNK UNK 六年级的男孩,是不是真的有助于避免。镜像和生活回忆的样子 UNK。

**Bi-LSTM (attention)**

而且她暗示我有些不正常，是因为我没有亲过。我的朋友们都没有男朋友。任何约会经历都是温迪 UNK，她是西雅图 UNK 学院唯一的一位学术卓越的女性，谁在 UNK 找到男性，它并没有帮助任何迫害。照片和记忆上看起来很难的生活。

**Deep Conv+LSTM**

她提到我有点不寻常，因为我还没有被吻过。我的朋友都没有男朋友。她是西雅图 UNK 学院唯一的学者，如果她在六年级找到一个男人，就不会受到迫害。镜子上的图片的记忆和生命看起来很棒。

**H-RNN Search**

她暗示我有点不正常，因为我还没有被吻过，这激怒了我。我的朋友还没有男朋友。西雅图优秀学院唯一有约会经历的女孩是温迪 UNK，她发现六年级的男生对她的拖延毫无帮助。摄影记忆和上镜看起来很艰难。

**HAN Encoder-decoder**

而她暗示我有点不正常，因为我没有被亲吻,这激怒了我。我的朋友们都没有男朋友。

西雅图学术卓越学院唯一一位有约会经历的女孩是温迪 UNK，她在六年级时发现了男孩-这当然没有帮助她拖延。摄影记忆和上镜看起来坚韧的生活。

**+ H-TRANS-joint**

她含沙射影地说我不正常，因为我还没有被吻过，这激怒了我。我的朋友们都没有男朋友。西雅图学术卓越学院唯一有约会经验的女孩是温迪 UNK，她在六年级就发现男孩-这对她的拖延毫无帮助。摄影记忆和照片相貌-艰难的生活。
latter two models Bi-LSTM (attention) and Deep Conv+LSTM incorrectly overlooking some contexts. In H-RNN Search, we also find some incorrect translations. And HAN encoder-decoder results in some incoherently translated clauses. In contrast, H-TRANS-joint is able to solve all these problems in translation to some extent.

5. Conclusion
To improve the performance of NMT in paragraph-level translation, we are the first to establish a paragraph parallel corpus and propose a hierarchical attention model. What distinguishes this study from previous ones is that a two-layer transformer model is applied in the encoder-decoder system to modify the input paragraph in a word-clause-paragraph structure, which allows the model to generate both inter-clause and intra-clause contexts in the paragraphs from source and target sides. As shown in the experimental results, our model significantly outperforms six competitive baselines on our model and TED Talks. The study shows the context from source and target can work in a complementary way to further improve translation performance. From the case study, it is found that our model can address most of the problems in paragraph translation to some extent.

In future work, we intend to explore the possibility of integrating an autoencoder into our model to get a better abstraction of the segmented clauses.

Acknowledgements
This research work has been funded by the National Natural Science Foundation of China (Grant No.61772337, U1736207), and the National Key Research and Development Program of China NO. 2018YFC0830703 and 2016QY03D0604.

References
[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need pp. 5998–6008 (2017)
[2] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the proper-ties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
[3] Li, J., Luong, M.T., Jurafsky, D.: A hierarchical neural autoencoder for paragraphs and documents. arXiv preprint arXiv:1506.01057 (2015)
[4] Bahdanau, Dzmitry, K.C., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
[5] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L., Mohanty, S.:Dividing and conquering long sentences in a translation system. In: Proceedings of the workshop on Speech and Natural Language. pp. 267–271. Association for Computational Linguistics (1992)
[6] Cettolo, M., Girardi, C., Federico, M.: Wit3: Web inventory of transcribed and translated talks. In: Conference of European Association for Machine Translation. pp. 261–268 (2012)
[7] Doi, T., Sumita, E.: Splitting input sentence for machine translation using language model with sentence similarity. In: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. pp. 113–119 (2004)
[8] Gong, Z., Zhang, M., Zhou, G.: Cache-based document-level statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 909–919. Association for Computational Linguistics(2011)
[9] Guillaume Lample, A.C.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
[10] Hardmeier, C., Federico, M.: Modelling pronominal anaphora in statistical machine translation. In: IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010. pp. 283–289 (2010)
[11] Jinsong Su, Jiali Zeng, D.X.Y.L.: A hierarchy-to-sequence attentional neural machine translation model. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING 26 (3), 623–632 (2018)
[12] Jiwei Li, M.T.L., Jurafsky, D.: Document-level neural machine translation with hierarchical attention networks[c]. EMNLP 2018 (2018), arXiv preprint arXiv:1809.01576., version 2
[13] Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810 (2017)
[14] Koehn, P., Hoang, H., Birch, A., Callison-Burch: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. pp. 177–180 (2007)
[15] Maruf, S., Haffari, G.: Document context neural machine translation with memory networks. arXiv preprint arXiv:1711.03688 (2017)
[16] Miculicich, L., Ram, D., Pappas: Document-level neural machine translation with hierarchical attention networks. arXiv preprint arXiv:1809.01576 (2018)
[17] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
[18] Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
[19] Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., Nie, J.Y.: A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 553–562. ACM (2015)
[20] Sudoh, K., Duh, K., Tsukada, H., Hirao, T., Nagata, M.: Divide and translate: improving long distance reordering in statistical machine translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. pp. 418–427. Association for Computational Linguistics (2010)
[21] Uszkoreit, J., Ponte, J.M., Popat, A.C., Dubiner, M.: Large scale parallel document mining for machine translation. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1101–1109. Association for Computational Linguistics (2010)
[22] Xu, J., Zens, R., Ney, H.: Sentence segmentation using ibm word alignment model 1. In: Proceedings of EAMT. pp. 280–287 (2005)
[23] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
[24] Zhou, J., Cao, Y., Wang, X., Li, P., Xu, W.: Deep recurrent models with fast-forward connections for neural machine translation. Transactions of the Association for Computational Linguistics 4, 371–383 (2016)
[25] Tu, Z., Liu, Y., Shi, S., Zhang, T.: Learning to remember translation history with a continuous cache. Transactions of the Association of Computational Linguistics 6, 407–420 (2018)