Document-Level Neural Machine Translation with Hierarchical Attention Networks

Lesly Miculicich†‡ Dhananjay Ram†‡ Nikolaos Pappas† James Henderson†
†Idiap Research Institute, Switzerland
‡École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{lmiculicich, dram, npappas, jhenderson}@idiap.ch

Abstract

Neural Machine Translation (NMT) can be improved by including document-level contextual information. For this purpose, we propose a hierarchical attention model to capture the context in a structured and dynamic manner. The model is integrated in the original NMT architecture as another level of abstraction, conditioning on the NMT model’s own previous hidden states. Experiments show that hierarchical attention significantly improves the BLEU score over a strong NMT baseline with the state-of-the-art in context-aware methods, and that both the encoder and decoder benefit from context in complementary ways.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017) trains an encoder-decoder network on sentence pairs to maximize the likelihood of predicting a target-language sentence given the corresponding source-language sentence, without considering the document context. By ignoring discourse connections between sentences and other valuable contextual information, this simplification potentially degrades the coherence and cohesion of a translated document (Hardmeier, 2012; Meyer and Webber, 2013; Sim Smith, 2017). Recent studies (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017; Tu et al., 2018) have demonstrated that adding contextual information to the NMT model improves the general translation performance, and more importantly, improves the coherence and cohesion of the translated text (Baden et al., 2018; Lapshinova-Koltunski and Hardmeier, 2017). Most of these methods use an additional encoder (Jean et al., 2017; Wang et al., 2017) to extract contextual information from previous source-side sentences. However, this requires additional parameters and it does not exploit the representations already learned by the NMT encoder. More recently, Tu et al. (2018) have shown that a cache-based memory network performs better than the above encoder-based methods. The cache-based memory keeps past context as a set of words, where each cell corresponds to one unique word keeping the hidden representations learned by the NMT while translating it. However, in this method, the word representations are stored irrespective of the sentences where they occur, and those vector representations are disconnected from the original NMT network.

We propose to use a hierarchical attention network (HAN) (Yang et al., 2016) to model the contextual information in a structured manner using word-level and sentence-level abstractions. In contrast to the hierarchical recurrent neural network (HRNN) used by (Wang et al., 2017), here the attention allows dynamic access to the context by selectively focusing on different sentences and words for each predicted word. In addition, we integrate two HANs in the NMT model to account for target and source context. The HAN encoder helps in the disambiguation of source-word representations, while the HAN decoder improves the target-side lexical cohesion and coherence. The integration is done by (i) re-using the hidden representations from both the encoder and decoder of previous sentence translations and (ii) providing input to both the encoder and decoder for the current translation. This integration method enables it to jointly optimize for multiple-sentences. Furthermore, we extend the original HAN with a multi-head attention (Vaswani et al., 2017) to capture different types of discourse phenomena.

Our main contributions are the following: (i) We propose a HAN framework for translation to capture context and inter-sentence connections in a structured and dynamic manner. (ii) We integrate the HAN in a very competitive NMT ar-
architecture (Vaswani et al., 2017) and show significant improvement over two strong baselines on multiple data sets. (iii) We perform an ablation study of the contribution of each HAN configuration, showing that contextual information obtained from source and target sides are complementary.

2 The Proposed Approach

The goal of NMT is to maximize the likelihood of a set of sentences in a target language represented as sequences of words \( y = (y_1, ..., y_t) \) given a set of input sentences in a source language \( x = (x_1, ..., x_m) \) as:

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

(1)

so, the translation of a document \( D \) is made by translating each of its sentences independently. In this study, we introduce dependencies on the previous sentences from the source and target sides:

\[
\text{max}_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log(P_{\Theta}(y^n|x^n, D_{xn}, D_{yn}))
\]

(2)

where \( D_{xn} = (x^{n-k}, ..., x^{n-1}) \) and \( D_{yn} = (y^{n-k}, ..., y^{n-1}) \) denote the previous \( k \) sentences from source and target sides respectively. The contexts \( D_{xn} \) and \( D_{yn} \) are modeled with HANs.

2.1 Hierarchical Attention Network

The proposed HAN has two levels of abstraction. The word-level abstraction summarizes information from each previous sentence \( j \) into a vector \( s^j \) as:

\[
q_w = f_w(h_t)
\]

(3)

\[
s^j = \text{MultiHead}(q_w, h^j_t)
\]

(4)

where \( h \) denotes a hidden state of the NMT network. In particular, \( h_t \) is the last hidden state of the word to be encoded, or decoded at time step \( t \), and \( h^j_t \) is the last hidden state of the \( j \)-th word of the \( i \)-th sentence of the context. The function \( f_w \) is a linear transformation to obtain the query \( q_w \). We used the MultiHead attention function proposed by (Vaswani et al., 2017) to capture different types of relations among words. It matches the query against each of the hidden representations \( h^j_t \) (used as value and key for the attention).

The sentence-level abstraction summarizes the contextual information required at time \( t \) in \( d_t \) as:

\[
q_s = f_s(h_t)
\]

(5)

\[
d_t = \text{FFN}(\text{MultiHead}(q_s, s^j))
\]

(6)

where \( f_s \) is a linear transformation, \( q_s \) is the query for the attention function, \( \text{FFN} \) is a position-wise feed-forward layer (Vaswani et al., 2017). Each layer is followed by a normalization layer (Lei Ba et al., 2016).

2.2 Context Gating

We use a gate (Tu et al., 2018, 2017) to regulate the information at sentence-level \( h_t \) and the contextual information at document-level \( d_t \). The intuition is that different words require different amount of context for translation:

\[
\lambda_t = \sigma(W_h h_t + W_dd_t)
\]

(7)

\[
\tilde{h}_t = \lambda_t h_t + (1 - \lambda_t) d_t
\]

(8)

where \( W_h, W_p \) are parameter matrices, and \( \tilde{h}_t \) is the final hidden representation for a word \( x_t \) or \( y_t \).

2.3 Integrated Model

The context can be used during encoding or decoding a word, and it can be taken from previously encoded source sentences, previously decoded target sentences, or from previous alignment vectors (i.e. context vectors (Bahdanau et al., 2015)). The different configurations will define the input query and values of the attention function. In this work we experiment with five of them: one at encoding time, three at decoding time, and one combining both. At encoding time the query is a function of the hidden state \( h_{xt} \) of the current word to be encoded \( x_t \), and the values are the encoded states of previous sentences \( h_{xt}^j \) (HAN encoder). At decoding time, the query is a function of the hidden state \( h_{yt} \) of the current word to be decoded \( y_t \), and the values can be (a) the encoded states of previous sentences \( h_{xt}^j \) (HAN decoder source),
(b) the decoded states of previous sentences \( h_{y_0} \)
(HAN decoder), and (c) the alignment vectors \( c_{ij} \)
(HAN decoder alignment). Finally, we combine complementary target-source sides of the context by joining HAN encoder and HAN decoder. Figure 1 shows the integration of the HAN encoder with the NMT model; a similar architecture is applied to the decoder. The output \( h_t \) is used by the NMT model as replacement of \( h_t \) during the final classification layer.

3 Experimental Setup

3.1 Datasets and Evaluation Metrics

We carry out experiments with Chinese-to-English (Zh-En) and Spanish-to-English (Es-En) sets on three different domains: talks, subtitles, and news.

TED Talks is part of the IWSLT 2014 and 2015 (Cettolo et al., 2012, 2015) evaluation campaigns. We use dev2010 for development; and tst2010-2012 (Es-En), tst2010-2013 (Zh-En) for testing. The Zh-En subtitles corpus is a compilation of TV subtitles designed for research on context (Wang et al., 2018). In contrast to the other sets, it has three references to compare. The Es-En corpus is a subset of OpenSubtitles2018 (Lison and Tiedemann, 2016). We randomly select two episodes for development and testing each. Finally, we use the Es-En News-Commentaries11 corpus which has document-level delimitation. We evaluate on WMT sets (Bojar et al., 2013): newstest2008 for development, and newstest2009-2013 for testing. A similar corpus for Zh-En is too small to be comparable. Table 2 shows the corpus statistics.

For evaluation, we use BLEU score (Papineni et al., 2002) on tokenized text, and we measure significance with the paired bootstrap resampling method proposed by Koehn (2004) (implementations by Koehn et al. (2007)).

3.2 Model Configuration and Training

As baselines, we use a NMT transformer, and a context-aware NMT transformer with cache memory which we implemented for comparison following the best model described by Tu et al. (2018), with memory size of 25 words. We used the OpenNMT (Klein et al., 2017) implementation of the transformer network. The configuration is the same as the model called “base model” in the original paper (Vaswani et al., 2017). The encoder and decoder are composed of 6 hidden layers each. All hidden states have dimension of 512, dropout of 0.1, and 8 heads for the multi-head attention. The target and source vocabulary size is 30K. The optimization and regularization methods were the same as proposed by Vaswani et al. (2017). Inspired by Tu et al. (2018) we trained the models in two stages. First we optimize the parameters for the NMT without the HAN, then we proceed to optimize the parameters of the whole network. We use \( k = 3 \) previous sentences, which gave the best performance on the development set.

4 Experimental Results

4.1 Translation Performance

Table 1 shows the BLEU scores for different models. The baseline NMT transformer already has better performance than previously published results on these datasets, and we replicate previous previous improvements from the cache method over this stronger baseline. All of our proposed HAN models perform at least as well as the cache method. The best scores are obtained by the combined encoder and decoder HAN model, which is significantly better than the cache method on all datasets without compromising training speed (2.3K vs 2.6K tok/sec). An important portion of the improvement comes from the HAN encoder, which can be attributed to the fact that the source-side always contains correct information, while the target-side may contain erroneous predictions at testing time. But combining HAN decoder with HAN encoder further improves translation performance, showing that they contribute complementary information. The three ways of incorporating information into the decoder all perform similarly.

Table 3 shows the performance of our best HAN model with a varying number \( k \) of previous sentences in the test-set. We can see that the best performance for TED talks and news is archived with 3, while for subtitles it is similar between 3 and 7.

4.2 Accuracy of Pronoun/Noun Translations

We evaluate coreference and anaphora using the reference-based metric: accuracy of pronoun translation (Miculicich Werlen and Popescu-Belis, 2017b), which can be extended for nouns. The list of evaluated pronouns is predefined in the metric, while the list of nouns was extracted using NLTK POS tagging (Bird, 2006). The upper part
sine similarity is calculated from one sentence to is used to obtain sentence representations, then co-

improvement in this aspect. Far from the human reference, so there is room for HAN boosts lexical cohesion, the scores are still not always make the translation better. Although as previously demonstrated in 4.2, repetitions do not.

Table 4 (bottom-left) displays the average ratio per tested document. In some cases, HAN decoder achieves the best score because it produces a larger quantity of repetitions than other models. However, as previously demonstrated in 4.2, repetitions do not always make the translation better. Although HAN boosts lexical cohesion, the scores are still far from the human reference, so there is room for improvement in this aspect.

For coherence, we use a metric based on Latent Semantic Analysis (LSA) (Foltz et al., 1998). LSA is used to obtain sentence representations, then cosine similarity is calculated from one sentence to

### Table 1: BLEU score for the different configurations of the HAN model, and two baselines. The highest score per dataset is marked in bold. Δ denotes the difference in BLEU score with respect of the NMT transformer. The significance values with respect to the NMT and the cache method are denoted by *, †, and ‡ respectively. The repetitions correspond to the p-values: * < .05, ** < .01, *** < .001.

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU</th>
<th>Δ</th>
<th>BLEU</th>
<th>Δ</th>
<th>BLEU</th>
<th>Δ</th>
<th>BLEU</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT transformer</td>
<td>16.87</td>
<td></td>
<td>35.44</td>
<td></td>
<td>28.60</td>
<td>35.20</td>
<td>21.36</td>
<td></td>
</tr>
<tr>
<td>+ cache (Tu et al., 2018)</td>
<td>17.32</td>
<td>(+0.45)**</td>
<td>36.46</td>
<td>(+1.02)**</td>
<td>28.86</td>
<td>(+0.26)</td>
<td>35.49</td>
<td>(+0.29)</td>
</tr>
<tr>
<td>+ HAN encoder</td>
<td>17.61</td>
<td>(+0.74)**</td>
<td>36.91</td>
<td>(+1.47)**</td>
<td>29.35</td>
<td>(+0.75)**</td>
<td>35.98</td>
<td>(+0.76)**</td>
</tr>
<tr>
<td>+ HAN decoder</td>
<td>17.39</td>
<td>(+0.52)**</td>
<td>37.01</td>
<td>(+1.57)**</td>
<td>29.21</td>
<td>(+0.61)</td>
<td>35.50</td>
<td>(+0.30)</td>
</tr>
<tr>
<td>+ HAN decoder source</td>
<td>17.56</td>
<td>(+0.69)**</td>
<td>36.94</td>
<td>(+1.50)**</td>
<td>28.92</td>
<td>(+0.32)</td>
<td>35.71</td>
<td>(+0.51)</td>
</tr>
<tr>
<td>+ HAN decoder alignment</td>
<td>17.48</td>
<td>(+0.61)**</td>
<td>37.03</td>
<td>(+1.60)**</td>
<td>28.87</td>
<td>(+0.27)</td>
<td>35.63</td>
<td>(+0.43)</td>
</tr>
<tr>
<td>+ HAN encoder + HAN decoder</td>
<td>17.79</td>
<td>(+0.92)**</td>
<td>37.24</td>
<td>(+1.80)**</td>
<td>29.67</td>
<td>(+1.07)**</td>
<td>36.23</td>
<td>(+1.03)**</td>
</tr>
</tbody>
</table>

Table 2: Dataset statistics in # sentence pairs.

<table>
<thead>
<tr>
<th>Models</th>
<th>TED Talks</th>
<th>Subtitles</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zh–En</td>
<td>Es–En</td>
<td>Zh–En</td>
</tr>
<tr>
<td>Training</td>
<td>0.2M</td>
<td>0.2M</td>
<td>2.2M</td>
</tr>
<tr>
<td>Development</td>
<td>0.8K</td>
<td>0.8K</td>
<td>1.1K</td>
</tr>
<tr>
<td>Test</td>
<td>5.5K</td>
<td>4.7K</td>
<td>1.2K</td>
</tr>
</tbody>
</table>

Table 3: Performance for variable context sizes k with the HAN encoder + HAN decoder.

Table 5 shows an example where HAN helped to generate the correct translation. The first box shows the current sentence with the analyzed word highlighted in red. The second box shows the past context at source and target. For the context visualization we use the toolkit provided by Pappas and Popescu-Belis (2017). Red corresponds to sentences, and blue to words. The intensity of color is proportional to the weight. We see that HAN correctly translates the ambiguous Spanish pronoun “él” into the English “his”. The HAN decoder highlighted a previous mention of “él”, and the HAN encoder highlighted the antecedent “Nathaniel”. This shows that HAN can capture interpretable inter-sentence connections. More samples with different attention heads are shown in the Appendix A.

### 5 Related Work

Statistical Machine Translation (SMT) Initial studies were based on cache memories (Tiede-

\[ \text{NIST BLEU} = \text{NMT transformer 35.99, cache 36.52, and HAN 37.15.} \]
### Table 4: Evaluation on discourse phenomena. Noun and pronoun translation: Accuracy with respect to a human reference. Lexical cohesion: Ratio of repeated and lexically similar words over the number of content words. Coherence: Average cosine similarity of consecutive sentences (i.e. average of LSA word-vectors)

<table>
<thead>
<tr>
<th>Model</th>
<th>TED Talks</th>
<th>Subtitles</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zh–En</td>
<td>Es–En</td>
<td>Zh–En</td>
</tr>
<tr>
<td>NMT Transformer</td>
<td>40.16</td>
<td>65.97</td>
<td>46.65</td>
</tr>
<tr>
<td>+ cache</td>
<td>40.87</td>
<td>66.75</td>
<td>46.00</td>
</tr>
<tr>
<td>+ HAN encoder</td>
<td>41.93</td>
<td>67.75</td>
<td>46.78</td>
</tr>
<tr>
<td>+ HAN decoder</td>
<td>41.61</td>
<td>67.35</td>
<td>46.78</td>
</tr>
<tr>
<td>+ HAN encoder + HAN decoder</td>
<td>42.99</td>
<td>67.81</td>
<td>47.43</td>
</tr>
<tr>
<td></td>
<td>62.30</td>
<td>50.40</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexical cohesion</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT Transformer</td>
<td>54.26</td>
</tr>
<tr>
<td>+ cache</td>
<td>54.37</td>
</tr>
<tr>
<td>+ HAN encoder</td>
<td>54.95</td>
</tr>
<tr>
<td>+ HAN decoder</td>
<td>55.40</td>
</tr>
<tr>
<td>+ HAN enc. + HAN dec.</td>
<td>56.08</td>
</tr>
<tr>
<td>Human reference</td>
<td>56.10</td>
</tr>
</tbody>
</table>

### Table 5: Example of pronoun disambiguation using HAN (TED Talks Es-En).

<table>
<thead>
<tr>
<th>Currently Translated Sentence</th>
<th>Context from Previous Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src.: y esto es un escape de su estado atormentado.</td>
<td>HAN decoder context with target. Query: his (En)</td>
</tr>
<tr>
<td>Ref.: and that is an escape from his tormented state.</td>
<td>a3^1 music is medicine . music changes us .</td>
</tr>
<tr>
<td>Base: and this is an escape from its &lt; unk &gt; state.</td>
<td>a3^2 and for Nathaniel , music is mine .</td>
</tr>
<tr>
<td>Cache: and this is an escape from their state.</td>
<td>because music allows him to take his thoughts and his delusions and turn through his imagination and his creativity actually .</td>
</tr>
<tr>
<td>HAN: and this is an escape from his &lt; unk &gt; state.</td>
<td>HAN encoder context with source. Query: su (Es)</td>
</tr>
<tr>
<td></td>
<td>a3^1 la música es medicina , la música nos cambia .</td>
</tr>
<tr>
<td></td>
<td>a3^2 y para Nathaniel la música es cordura .</td>
</tr>
<tr>
<td></td>
<td>because the music lets you transform your thoughts and your delusions and transform them to his imagination and his creativity .</td>
</tr>
<tr>
<td></td>
<td>n y su creatividad en realidad .</td>
</tr>
</tbody>
</table>

mann, 2010; Gong et al., 2011). However, most of the work explicitly models discourse phenomena (Sim Smith, 2017) such as lexical cohesion (Meyer and Popescu-Belis, 2012; Xiong et al., 2013; Loáiciga and Grisot, 2016; Pu et al., 2017; Mascarell, 2017), coherence (Born et al., 2017), and coreference (Rios Gonzales and Tuggener, 2017; Miculicich Werlen and Popescu-Belis, 2017a). Hardmeier et al. (2013) introduced the document-level SMT paradigm.

### Sentence-level NMT

Initial studies on NMT enhanced the sentence-level context by using memory networks (Wang et al., 2016), self-attention (Miculicich Werlen et al., 2018; Zhang et al., 2016), and latent variables (Yang et al., 2017).

### Document-level NMT


### 6 Conclusion

We proposed a hierarchical multi-head HAN NMT model\(^5\) to capture inter-sentence connections. We integrated context from source and target sides by directly connecting representations from previous sentence translations into the current sentence translation. The model significantly outperforms two competitive baselines, and the ablation study shows that target and source context is complementary. It also improves lexical cohesion and coherence, and the translation of nouns and pronouns. The qualitative analysis shows that the model is able to identify important previous sentences and words for the correct prediction. In future work, we plan to explicitly model discourse connections with the help of annotated data, which may further improve translation quality.

### Acknowledgments

We are grateful for the support of the European Union under the Horizon 2020 SUMMA project n. 688139, see [www.summa-project.eu](http://www.summa-project.eu).

---

References


