EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

Organised by

Universitat d'Alacant

Universidad de Alicante

transducens

research group

ISBN: 978-84-09-01901-4

Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides

Ruchit Agrawal[1,2], **Marco Turchi**[1], **Matteo Negri**[1]
[1]Fondazione Bruno Kessler, Italy
[2]University of Trento, Italy
{ragrawal, turchi, negri}@fbk.eu

Abstract

A salient feature of Neural Machine Translation (NMT) is the end-to-end nature of training employed, eschewing the need of separate components to model different linguistic phenomena. Rather, an NMT model learns to translate individual sentences from the labeled data itself. However, traditional NMT methods trained on large parallel corpora with a one-to-one sentence mapping make an implicit assumption of sentence independence. This makes it challenging for current NMT systems to model inter-sentential discourse phenomena. While recent research in this direction mainly leverages a single previous source sentence to model discourse, this paper proposes the incorporation of a context window spanning previous as well as next sentences as source-side context and previously generated output as target-side context, using an effective non-recurrent architecture based on self-attention. Experiments show improvement over non-contextual models as well as contextual methods using only previous context.

1 Introduction

Neural Machine Translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014) has consistently outperformed other MT paradigms across a range of domains, applications and training settings (Bentivogli et al., 2016; Castilho et al., 2017; Toral and Sánchez-Cartagena, 2017), thereby emerging as the *de facto* standard in Machine Translation. NMT models are typically trained at the sentence level (Cho et al., 2014), whereby the probability of an output sentence given an input sentence is maximized, implicitly making an assumption of sentence independence across the dataset. This works well for the translation of stand-alone sentences or datasets containing shuffled sentences, which are not connected with each other in terms of discursive dependencies. However, in real life situations, written text generally follows a sequential order featuring a number of cross-sentential phenomena. Additionally, speech-like texts (Bawden, 2017) exhibit the trait of contextual dependency and sequentiality as well, often containing a greater number of references that require a common knowledge ground and discourse understanding for correct interpretation. Figure 1 shows an example of such inter-sentential dependencies. These dependencies are not fully leveraged by the majority of contemporary NMT models, owing to the treatment of sentences as independent units for translation.

In order to perform well on sequential texts, NMT models need access to extra information, which could serve as the disambiguating context for better translation. Recent work in this direction (Zoph and Knight, 2016; Jean et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2017; Wang et al., 2017) has primarily focused on previous source-side context for disambiguation. Since all of these approaches utilize recurrent architectures, adding context comprising of more than a single previous sentence can be challenging due to either (i) the increased number of estimated parameters and training time, in case of the multi-encoder approach (Jean et al., 2017), or (ii)

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 11–20
Alacant, Spain, May 2018.

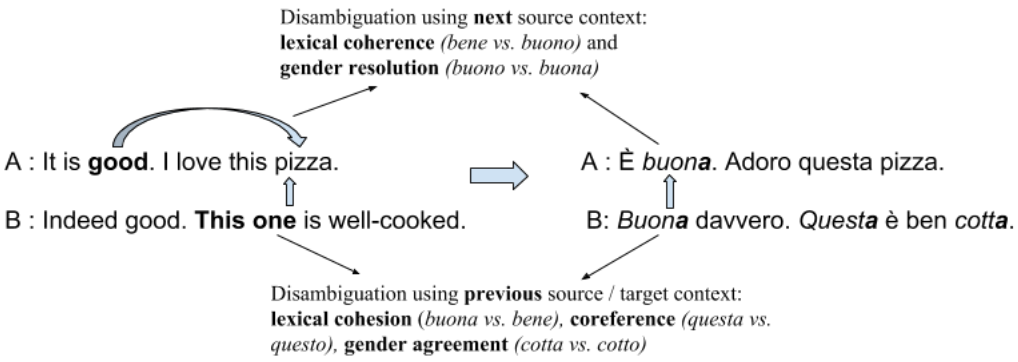

Figure 1: Inter-sentential dependencies requiring previous (source and target) and next (source) context

performance drop due to very long inputs (Koehn and Knowles, 2017), in case of extended translation units (Tiedemann and Scherrer, 2017). Hence, the impact of utilizing a large-sized context window on the source as well as the target side remains unclear. Additionally, the impact of incorporating the next sentences as context in the source side also needs to be examined, owing to discourse phenomena like cataphora and gender agreement, illustrated in Figure 1.

We address this gap and investigate the contribution of a context window looking behind as well as ahead on the source-side, combined with previous target-side context, in an efficient non-recurrent "Transformer" architecture with self-attention (*hereafter Transformer*), recently proposed by Vaswani et al. (2017). We choose this architecture due to its effective handling of long-range dependencies and easily achievable computational parallelization. These characteristics are due to the fact that the Transformer is based entirely on self-attention, as opposed to LSTMs or GRUs. The non-recurrent architecture enables effective parallelization, which is not possible with RNNs due to their sequentiality, thereby reducing the computational complexity considerably. We perform experiments using differently sized context windows on the source and target side. This is the first effort towards contextual NMT with Transformer to the best of our knowledge. On the English-Italian data from the IWSLT 2017 shared task (Cettolo et al., 2017), the best of our models achieves a 2.3% increase in BLEU score over a baseline Transformer model trained without any inter-sentential context and a 2.6% increase in BLEU score over a multi-source BiLSTM model trained using a previous source sentence as additional context.

The major contributions of this paper are summarized below:

- We demonstrate that looking ahead at the following text in addition to looking behind at the preceding text on the source-side improves performance.
- We demonstrate that both source-side context as well as target-side context help to improve translation quality, the latter however is more prone to error propagation.
- We demonstrate that looking further beyond a single previous sentence on the source-side results in better performance, especially in absence of target-side context.
- We show that a simple method like concatenation of the multiple inputs, when used with the Transformer, generates efficient translations, whilst being trained more than three times faster than an RNN based architecture.

The rest of the paper is organized as follows: We describe an outline of the related work in Section 2, and provide a theoretical background in Section 3. Section 4.1 briefly describes the discourse phenomena which we would like to capture using our contextual NMT models. Our approach to model discourse and the experiments conducted are described in Section 4. Section Section 5 presents the results obtained by our models, along with a linguistic analysis of the implications therein. We present the conclusions of the present research and highlight possible directions for future work in Section 6.

2 Related Work

Discourse modeling has been explored to a significant extent for Statistical Machine Translation (Hardmeier, 2012), using methods like discriminative learning (Giménez and Màrquez, 2007; Tamchyna et al., 2016), context features (Gimpel and Smith, 2008; Costa-Jussà et al., 2014; Sánchez-Martínez et al., 2008; Vintar et al., 2003), bilingual language models (Niehues et al., 2011), document-wide decoding (Hardmeier et al., 2012; Hardmeier et al., 2013) and factored models (Meyer et al., 2012). The majority of these works, however, look mainly at intra-sentential discourse phenomena, owing to the limited capability of SMT models to exploit extra-sentential context. The neural MT paradigm, on the other hand, offers a larger number of avenues for looking beyond the current sentence during translation.

Recent work on incorporating contextual information into NMT models has delved primarily into multi-encoder models (Zoph and Knight, 2016; Jean et al., 2017; Bawden et al., 2017), hierarchy of RNNs (Wang et al., 2017) and extended translation units containing the previous sentence (Tiedemann and Scherrer, 2017). These approaches build upon the multi-task learning method proposed by Luong et al. (2015), adapting it specifically for translation. Zoph and Knight (2016) propose a multi-source training method, which employs multiple encoders to represent inputs coming from different languages. Their method utilizes the sources available in two languages in order to produce better translations for a third language. Jean et al. (2017) use the multi-encoder framework, with one set of encoder and attention each for the previous and the current source sentence as an attempt to model context. However, this method would be computationally expensive with an increase in the number of contextual sentences owing to the increase in estimated parameters.

Wang et al. (2017) employ a hierarchy of RNNs to summarize source-side context (previous three sentences). This method addresses the computational complexity to an extent, however it does not incorporate target-side context, which has been shown to be useful by (Bawden et al., 2017). Bawden et al. (2017) present an in-depth analysis of the evaluation of discourse phenomena in NMT and the challenges faced thereof. They provide a hand-crafted test set specifically aimed at capturing discursive dependencies. However, this set is created with the assumption that the disambiguating context lies in the previous sentence, which is not always the case (Scarton et al., 2015).

Our work is most similar to (Tiedemann and Scherrer, 2017), who employ the standard NMT architecture without multiple encoders, but using larger blocks containing the previous and the current sentence as input for the encoder, as an attempt to better model discourse phenomena. The primary limitation of this method is the inability to add larger context due to the ineffective handling of long-range dependencies by RNNs (Koehn and Knowles, 2017). Additionally, this method does not look at the following source-text, due to which phenomena like cataphora and lexical cohesion are not captured well.

While the above-mentioned works employ the previous source text, we propose employing a context window spanning previous as well as next source sentences in order to model maximal discourse phenomena. On the target-side, we decode the previous and current sentence while looking at the source-window, thereby employing target-side context as well. Additionally, we employ the Transformer for our contextual models, as opposed to the above-mentioned works using RNNs, due to the enhanced long-range performance and computational parallelization.

3 Background

3.1 NMT with RNNs and Transformer

Neural MT employs a single neural network trained jointly to provide end-to-end translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014). NMT models typically consist of two components - an encoder and a decoder. The components are generally composed of Stacked RNNs (Recurrent Neural Networks), using either Long Short Term Memory (LSTM) (Sundermeyer et al., 2012) or Gated Recurrent Units (GRU) (Chung et al., 2015). The encoder transforms the source sentence into a vector from which the decoder extracts the probable targets. Specifically, NMT aims to model the conditional probability $p(y|x)$ of translating a source sentence x = $x_1, x_2...x_u$ to a target sentence y = $y_1, y_2, ...y_v$. Let s be the representation of the source sentence as computed by the encoder. Based on the source representation, the decoder produces a translation, one target word at a time

and decomposes the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^{v} \log p(y_j|y_{<j}, s) \quad (1)$$

The entire model is jointly trained to maximize the (conditional) log-likelihood of the parallel training corpus:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y^{(n)}|x^{(n)}, \theta) \quad (2)$$

where $(y^{(n)}, x^{(n)})$ represents the n^{th} sentence in parallel corpus of size N and θ denotes the set of all tunable parameters.

Research in NMT recently witnessed a major breakthrough in the Transformer architecture proposed by Vaswani et al. (2017). This architecture eschews the recurrent as well as convolution layers, both of which are integral to the majority of contemporary neural network architectures. Instead, it uses stacked multi-head attention as well as positional encodings to model the complete sequential information encoded by the input sentences. The decoder comprises of a similar architecture, using masked multi-head attention followed by softmax normalization to generate the output probabilities over the target vocabulary. The positional encodings are added to the input as well as output embeddings, enabling the model to capture the sequentiality of the input sentence without having recurrence. The encodings are computed from the position (pos) and the dimension (i) as follows:

$$PE_{(pos,2i)} = sin(pos/10000^{(2i/d_{model})}) \quad (3)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{(2i/d_{model})}) \quad (4)$$

where PE stands for positional encodings and d_{model} is the dimensionality of the vectors resulting from the embeddings learned from the input and output tokens. Thus, each dimension of the encoding (i) corresponds to a sinusoid.

3.2 Inter-sentential discourse phenomena

Coherence in a text is implicitly established using a variety of discourse relations. Contextual information can help in handling a variety of discourse phenomena, mainly involving lexical choice, linguistic agreement, coreference - anaphora (Hardmeier and Federico, 2010) as well as cataphora, and lexical coherence. Spoken language especially contains a large number of such dependencies, due to the presence of an environment facilitating direct communication between the parties (Pierrehumbert and Hirschberg, 1990), where gestures and a common ground/theme are often used the disambiguating context, thereby rendering the need for explicit mentions in the text less important. A reasonable amount of noun phrases are established deictically, and the theme persists until it's taken over by another theme.

The deictic references are challenging to resolve for NMT models using only the current sentence-pair in consideration, and possible errors involving gender usage as well as linguistic agreement can be introduced in the translation. For instance, for English → Italian translation, establishing the linguistic features of the noun under consideration is crucial for translation. The co-ordination with the adjective (*buona* vs *buono*), pronominal references (*lui* vs *lei*), past participle verb form (*sei andato* vs *sei andata*) as well as articles (*il* vs *la*) depends on the noun.

Establishing the noun under consideration could improve MT quality significantly, an example of which is shown in (Babych and Hartley, 2003), wherein Named Entity Recognition benefit translation. This would eventually lead to less post-editing effort, which is significant for correcting coreference related errors (Daems et al., 2015). Other inter-sentential phenomena we would like to capture include temporality (precedence, succession), causality (reason, result), condition (hypothetical, general, unreal, factual), implicit assertion, contrast (juxtaposition, opposition) and expansion (conjunction, instantiation, restatement, alternative).

4 Experiments

4.1 Context integration

We model discourse using context windows on the source as well as the target side. For the source, we use one, two and three previous sentences and one next sentence as additional context. For the target, we use one and two previous sentences as additional context.[1] We choose the Transformer for our experiments. The non-recurrent architecture enables it to better handle longer sequences, without an additional computational cost. This

[1] Increasing beyond this caused a drop in performance in our preliminary experiments.

is made possible by using a multi-headed self-attention mechanism. The attention is a mapping from (query, key, value) tuples to an output vector. For the self-attention, the query, key and value come from the previous encoder layer, and the attention is computed as:

$$SA(Q, K, V) = softmax(QK^T/\sqrt{d_k})V \quad (5)$$

where Q is the query matrix, K is the key matrix and V is the value matrix, d_k is the dimensionality of the queries and keys, and SA is the computed self-attention. This formulation ensures that the net path length between any two tokens irrespective of their position in the sequence is O(1).

The multi-head attention makes it possible for the Transformer to model information coming in from different positions simultaneously. It employs multiple attention layers in parallel, with each head using different linear transformations and thereby learning different relationships, to compute the net attention:

$$MH(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (6)$$

where MH is the multi-head attention, h is the number of attention layers (also called "heads"), $head_i$ is the self-attention computed over the i^{th} attention layer and W^O is the parameter matrix of dimension hd_v*d_{model}. In this case, queries come from the previous decoder layer, and the key-value pairs come from encoder output.

For training the contextual models, we investigate the usage of all the possible combinations from the following configurations for modeling context on both sides:

- Source side configuration:
  - Previous sentence, previous two sentences, previous three sentences, previous and next sentence, previous two and next sentence.
- Target side configuration:
  - Previous sentence, previous two sentences.

For our experiments using the Transformer model, we concatenate the contextual information in our training and validation sets using a *BREAK* token, inspired by (Tiedemann and Scherrer, 2017). Since the Transformer has positional encodings, it encodes position information inherently and using just a single *BREAK* token worked better than appending a feature for each token specifying the sentence it belongs to. The models are referred to by the following label subsequently:

$$Prev_m + Curr + Next_n \rightarrow Prev_p + Curr$$

where *m* is the number of previous sentences used as source-side context, *n* is the number of next sentences used as source-side context, and *p* is the number of previous sentences used as target-side context. $Curr$ refers to the current sentence on both sides.

For comparison with RNN based techniques, we trained baseline as well as contextual models using a BiLSTM architecture. We employed the previous sentence as source-side context for the contextual models, integrated using the methods of concatenation and multi-encoder RNN proposed by Tiedemann and Scherrer (2017) and Jean et al. (2017) respectively. These are denoted by the labels $concat$ and $Multi-Source$. For the concatenation, the $BREAK$ token was used, similar to the Transformer experiments. We also compared the performance using target-side context (Tiedemann and Scherrer, 2017; Bawden et al., 2017). The contextual models using only source-context are labeled "2 to 1", while those using the previous target sentence as context are labeled "2 to 2".

4.2 Dataset

For our experiments, we employ the IWSLT 2017 (Cettolo et al., 2012) dataset, for the language direction English → Italian (en → it). The dataset contains parallel transcripts of around 1000 TED talks, spanning various genres like Technology, Entertainment, Business, Design and Global issues.[2] We use the "train" set for training, the "tst2010" set for validation, and the "tst2017" set for testing. The statistics for the training, validation and test splits are as given in Table 1. For training the models, the sentences are first tokenized, following by segmentation of the tokens into subword units (Sennrich et al., 2015) using Byte Pair Encoding (BPE). The number of BPE operations is set to 32,000 and the frequency threshold for the vocabulary filter is set to 35.

[2]This dataset is publicly available at https://wit3.fbk.eu/

Phase	Training	Validation	Test
#Sentences	221,688	1,501	1,147
#Tokens-en	4,073,526	27,191	21,507
#Tokens-it	3,799,385	25,131	20,238

Table 1: Statistics for the IWSLT dataset

4.3 Model Settings

We employ OpenNMT-tf (Klein et al., 2017) for all our experiments.[3] For training the Transformer models, we use the Lazy Adam optimizer, with a learning rate of 2.0 , model dimension of 512, label smoothing of 0.1, beam width of 4, batch size of 3,072 tokens, bucket width of 1 and stopping criteria at 250,000 steps or plateau in BLEU, in case of the larger context models, since we observed some instability in the convergence behavior of the Transformer, especially for the contextual models. The maximum source length is set to be 70 for the baseline model, increasing linearly with more context. The maximum target length is set to be 10% more than the source length.[4] For training the RNN models, we employ the stochastic gradient descent optimizer, with a learning rate of 1.0, decay rate 0.7 with an exponential decay, beam width of 5, batch size 64, bucket width 1 and stopping criteria 250,000 steps or plateau in BLEU, whichever occurs earlier.

4.4 Evaluation

The evaluation of discourse phenomena in MT is a challenging task (Hovy et al., 2002; Carpuat and Simard, 2012), requiring specialized test sets to quantitatively measure the performance of the models for specific linguistic phenomena. One such test set was created by (Bawden et al., 2017) to measure performance on coreference, cohesion and coherence respectively. However, the test set was created with the assumption that the disambiguating context always lies in the previous sentence, which is not necessarily the case. Traditional automatic evaluation metrics do not capture discourse phenomena completely (Scarton et al., 2015), and using information about the discourse structure of a text improves the quality of MT evaluation (Guzmán et al., 2014). Hence, alternate methods for evaluation have been proposed (Mitkov et al., 2000; Fomicheva and Bel, 2016) However, these methods do not look at the document as a whole, but mainly model intra-sentential discourse. Developing an evaluation metric that considers document-level discourse remains an open problem. Hence, we perform a preliminary qualitative analysis in addition to the automatic evaluation of our outputs.

For automatic evaluation, we measure the performance of our models using two standard metrics: BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). For comparison with the test set, we extract the current sentence separated by the *BREAK* tokens from the output generated by the contextual models. We also measure the percentage of sentences for which the contextual models improve over the baseline model. This is done by computing the sentence-level TER for each generated output sentence, and comparing it with the corresponding one in the test set.

[3]The code is publicly available at https://github.com/OpenNMT/OpenNMT-tf

[4]This is done to ensure no loss in target-side information, a known sensitivity of the Transformer architecture.

Configuration	BLEU	TER
(i) BiLSTM, no context	28.2	52.9
(ii) BiLSTM, Concat, 2 to 1	26.3	53.7
(iii) BiLSTM, Multi-Source, 2 to 1	28.9	52.6
(iv) BiLSTM, Concat, 2 to 2	25.4	53.4
(v) BiLSTM, Multi-Source, 2 to 2	28.9	52.5

Table 2: Performance using RNN based approaches

Model Configuration	BLEU	TER
(i) $Curr \rightarrow Curr$	29.2	52.8
(ii) $Prev_1$ + $Curr \rightarrow Curr$	29.4	52.5
(iii) $Prev_2$ + $Curr \rightarrow Curr$	29.8	51.9
(iv) $Prev_3$ + $Curr \rightarrow Curr$	29.2	52.8
(v) $Curr$ + $Next_1 \rightarrow Curr$	29.7	51.9
(vi) $Prev_1$ + $Curr$ + $Next_1 \rightarrow Curr$	30.6	51.1
(vii) $Prev_2$ + $Curr$ + $Next_1 \rightarrow Curr$	29.8	51.4

Table 3: Results of our models using only source-side context, on en $\rightarrow$ it, IWSLT 2017

5 Results and Discussion

5.1 Performance on automatic evaluation metrics

Tables 3 and 4 show the results obtained by the different configurations of our models using the Transformer architecture. For comparison with previous approaches, we also train four contextual configurations using RNN-based models, and report the results in Table 2.

The RNN results confirm that:

- Adding contextual information is useful for RNN models, provided that it is incorporated using a multi-encoder architecture ($\approx$ 28.9

Model Configuration	BLEU	TER
(i) $Prev_1$ + $Curr \rightarrow Prev_1$ + $Curr$	29.5	52.1
(ii) $Prev_2$ + $Curr \rightarrow Prev_1$ + $Curr$	29.8	51.9
(iii) $Prev_2$ + $Curr \rightarrow Prev_2$ + $Curr$	29.7	52.1
(iv) $Prev_3$ + $Curr \rightarrow Prev_1$ + $Curr$	29.2	52.2
(v) $Prev_3$ + $Curr \rightarrow Prev_2$ + $Curr$	28.9	52.9
(vi) $Prev_1$ + $Curr$ + $Next_1 \rightarrow Prev_1$ + $Curr$	31.5	49.7
(vii) $Prev_2$ + $Curr$ + $Next_1 \rightarrow Prev_1$ + $Curr$	31.1	50.5
(viii) $Prev_2$ + $Curr$ + $Next_1 \rightarrow Prev_2$ + $Curr$	30.2	51.2

Table 4: Results of our models using source as well as target side context, on en → it, IWSLT 2017

Model Configuration	% sentences
$Prev_1$ + $Curr \rightarrow Curr$	62.8
$Curr$ + $Next_1 \rightarrow Curr$	61.3
$Prev_1$ + $Curr$ + $Next_1 \rightarrow Curr$	67.2

Table 5: Percentage of sentences for which TER score is less than or equal to the baseline model, depending upon the source-context used

BLEU score with multi-source, ≈ 0.8 more than the baseline BLEU score of 28.18).

- RNNs are sensitive to the length of the sentence, both on the source and target side (Table 2, (ii) and (iv)). This can be attributed to a vanishing signal between very long-range dependencies, despite the gating techniques employed.

- The RNN models need more sophisticated techniques than concatenation, like multi-source training, to leverage the information from the previous sentence (Table 2, (iii), (v)). This can be attributed to the drop in performance on very long sequences (Cho et al., 2014; Koehn and Knowles, 2017)[5], owing to concatenation.

For the Transformer architecture, the contextual models achieve an increase of 1-2% in BLEU score over a baseline model trained without any inter-sentential context (Tables 3 and 4).

The results suggest that:

- Looking further ahead at the next sentence can help in disambiguation, evident from the improved performance of the configurations involving both previous as well as next sentences on the source side than those looking only at previous context (Table 3, (v) - (vii)).

- Target-side context also helps to improve performance (Table 4, (i)-(v) vs. Table 3. (ii)-(iv)). as also suggested by (Bawden et al., 2017). However, a larger context window on the source side and a window with one previous sentence on the target side generally works better. Our intuition is that going beyond one previous sentence on the target side increases the risk of error propagation (Table 4, (viii)).

- The Transformer performs significantly better than RNN's for very long inputs (Table 2, (iv) vs. Table 4, (i)). This can be attributed to the multi-head self-attention, which captures long-range dependencies better.

- Contextual information does not necessarily come from the previous one sentence. Incorporating more context, especially on source-side, helps on TED data (Table 4, (vi), (vii)), and can be effectively handled with Transformer.

- The self-attention mechanism of the Transformer architecture enables a simple strategy like concatenation of a context window to work better than multi-encoder RNN based approaches.

Additionally, the training time for the Transformer models was significantly shorter than the RNN based ones (≈ 30 hours and ≈ 100 hours respectively). This can be attributed to the fact that the positional encodings capture the sequentiality in the absence of recurrence, and the multi-head attention makes it easily parallelizable. In addition to the corpus level scores, we also compute sentence level TER scores, in order to estimate the percentage of sentences which are better translated using cross sentential source-side context. These are given in Table 5.

5.2 Qualitative analysis

In addition to the performance evaluation using the automatic evaluation metrics, we also analyzed a random sample of outputs generated by our models, in order to have a better insight as to which linguistic phenomena are handled better by our contextual NMT models. Tables 6 and 7 compare the outputs of our best-performing contextual models (Table 4, (vi)) with the baseline model. The contextual models in general make better morphosyntactic choices generating more coherent translations than the baseline model. For instance, in the output of the contextual model (Table 6, (iii)), the

[5]On manual inspection, we observed frequent short, incomplete predictions in this case.

Source	I went there with **my friend**. She was amazed to see that it had multiple floors. **Each one** had a number of shops.
(i) Baseline Transformer	Arrivai li con **il mio amico**. Rimaneva meravigliato di vedere che aveva una cosa piu incredibile. **Ognuna** aveva tanti negozi.
(ii) Contextual Transformer (Prev)	Arrivai la con **il mio amico**. Era sorpresa vedere che aveva diversi piani. **Ognuno** aveva un certo numero di negozi.
(iii) Contextual Transformer (Prev + Next)	Sono andato con **la mia amica**. Fu sorpresa nel vedere che aveva piu piani. **Ognuno** aveva tanti negozi.
Reference	Sono andato la' con la mia amica. E' rimasta meraviglia nel vedere che aveva piu' piani. Ognuno aveva tanti negozi.

Table 6: Qualitative analysis - Improvement for cataphora, anaphora and gender agreement

Source	OK, I need you to take out your phones. Now that you have your phone out, I'd like you to unlock your phone.
(i) Baseline Transformer	Ok, devo **tirare** fuori i vostri cellulari. Ora che avete il vostro telefono, vorrei che bloccaste il vostro telefono.
(ii) Contextual Transformer (Prev)	OK, dovete tirare i vostri **cellulari**. Ora che avete il vostro telefono, vorrei che faceste sbloccare il vostro telefono.
(iii) Contextual Transformer (Prev + Next)	Ok, ho bisogno che **tiriate** fuori i vostri **telefoni**. Ora che avete il vostro telefono, vorrei che sbloccaste il vostro telefono.
Reference	Ok, ho bisogno che tiriate fuori i vostri telefoni. Ora che avete il vostro telefono davanti vorrei che lo sbloccaste.

Table 7: Qualitative analysis - Improvement for lexical cohesion and verbal inflections

phrase *sono andato* employs the *passato prossimo* ("near past") verb form *andato*, which is more appropriate than the *passato remoto* ("remote past") form *arrivai*, since the latter refers to events occurred far in the past, while the former refers to more recent ones. Additionally, the cataphor *my friend* is successfully disambiguated to refer to the postcedent *she*, apparent from the correctly predicted gender of the translated phrase *la mia amica* (feminine) as opposed to *il mio amico* (masculine). Similarly, the anaphora *Each one* is resolved (*ognuna* as opposed to *ognuno*). In the second example from Table 7, improved lexical choice *-che tiriate* (second person plural subjunctive), *bisogno* ("I need") as opposed to *devo* ("I must") and lexical cohesion *cellulari* ("mobile phones") *vs. telefoni* ("phones") can be observed.

While our models are able to incorporate contextual information from the surrounding text, they cannot leverage the disambiguating context which lies very far away from the current sentence being translated. In such cases, concatenating the sentences would be non-optimal, since there is a high possibility of irrelevant information overpowering disambiguating context. This is also evident from our experiments using $n > 2$ previous sentences as additional context using concatenation (Table 3, (iv)).

6 Conclusion

Neural MT methods, being typically trained at sentence level, fail to completely capture implicit discourse relations established at the inter-sentential level in the text. In this paper, we demonstrated that looking behind as well as peeking ahead in the source text during translation leads to better performance than translating sentences in isolation. Additionally, jointly decoding the previous as well as current text on the target-side helps to incorporate target-side context, which also shows improvement in translation quality to a certain extent, albeit being more prone to error propagation with increase in the size of the context window. Moreover we showed that using the Transformer architecture, a simple strategy like concatenation of the context yields better performance on spoken texts than non-contextual models, whilst being trained significantly faster than recurrent architectures. Contextual handling using self-attention is hence a promising direction to explore in the future, possibly with multi-source techniques in conjugation with the Transformer architecture. In the future, we would like to perform a fine-grained analysis on the improvement observed for specific linguistic phenomena using our extended context models.

References

Babych, Bogdan and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 1–8. Association for Computational Linguistics.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Bawden, Rachel. 2017. Machine translation of speech-like texts: Strategies for the inclusion of context. In *19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Carpuat, Marine and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449. Association for Computational Linguistics.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of pbsmt and nmt using professional translators.

Cettolo, Mauro, Girardi Christian, and Federico Marcello. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Cettolo, Mauro, Federico Marcello, Bentivogli Luisa, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chung, Junyoung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075.

Costa-Jussà, Marta R, Parth Gupta, Paolo Rosso, and Rafael E Banchs. 2014. English-to-hindi system description for wmt 2014: deep source-context features for moses. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 79–83.

Daems, Joke, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2015. The impact of machine translation error types on post-editing effort indicators. In *4th Workshop on Post-Editing Technology and Practice (WPTP4)*, pages 31–45. Association for Machine Translation in the Americas.

Fomicheva, Marina and Núria Bel. 2016. Using contextual information for machine translation evaluation.

Giménez, Jesús and Lluís Màrquez. 2007. Context-aware discriminative phrase selection for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 159–166. Association for Computational Linguistics.

Gimpel, Kevin and Noah A Smith. 2008. Rich source-side context for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17. Association for Computational Linguistics.

Guzmán, Francisco, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 687–698.

Hardmeier, Christian and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Hardmeier, Christian, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190. Association for Computational Linguistics.

Hardmeier, Christian, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics); 4-9 August 2013; Sofia, Bulgaria*, pages 193–198. Association for Computational Linguistics.

Hardmeier, Christian. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75.

Jean, Sebastien, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Kalchbrenner, Nal and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Luong, Minh-Thang, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Meyer, Thomas, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, number EPFL-CONF-192524.

Mitkov, Ruslan, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones, and Violeta Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49–58. Citeseer.

Niehues, Jan, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider context by using bilingual language models in machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 198–206. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pierrehumbert, Janet and Julia Bell Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*, pages 271–311.

Sánchez-Martínez, Felipe, Juan Antonio Pérez-Ortiz, and Mikel L Forcada. 2008. Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66.

Scarton, Carolina, Marcos Zampieri, Mihaela Vela, Josef van Genabith, and Lucia Specia. 2015. Searching for context: a study on document-level labels for translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tamchyna, Aleš, Alexander Fraser, Ondřej Bojar, and Marcin Junczys-Dowmunt. 2016. Target-side context for discriminative models in statistical machine translation. *arXiv preprint arXiv:1607.01149*.

Tiedemann, Jörg and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Toral, Antonio and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Vintar, Špela, Ljupčo Todorovski, Daniel Sonntag, and Paul Buitelaar. 2003. Evaluating context features for medical relation mining. *Data Mining and Text Mining for Bioinformatics*, page 64.

Wang, Longyue, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *arXiv preprint arXiv:1704.04347*.

Zoph, Barret and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.