Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations

Sameen Maruf[1]
Monash University
VIC, Australia

André F. T. Martins[2]
Unbabel &
Instituto de Telecomunicacões
Lisbon, Portugal

Gholamreza Haffari[1]
Monash University
VIC, Australia

[1]{firstname.lastname}@monash.edu
[2]andre.martins@unbabel.com

Abstract

Recent works in neural machine translation have begun to explore document translation. However, translating online multi-speaker conversations is still an open problem. In this work, we propose the task of translating Bilingual Multi-Speaker Conversations, and explore neural architectures which exploit both source and target-side conversation histories for this task. To initiate an evaluation for this task, we introduce datasets extracted from Europarl v7 and OpenSubtitles2016. Our experiments on four language-pairs confirm the significance of leveraging conversation history, both in terms of BLEU and manual evaluation.

1 Introduction

Translating a conversation online is ubiquitous in real life, e.g. in the European Parliament, United Nations, and customer service chats. This scenario involves leveraging the conversation history in multiple languages. The goal of this paper is to propose and explore a simplified version of such a setting, referred to as Bilingual Multi-Speaker Machine Translation (Bi-MSMT), where speakers' turns in the conversation switch the source and target languages. We investigate neural architectures that exploit the bilingual conversation history for this scenario, which is a challenging problem as the history consists of utterances in both languages.

The ultimate aim of all machine translation systems for dialogue is to enable a multi-lingual conversation between multiple speakers. However, translation of such conversations is not well-explored in the literature. Recently, there has been work focusing on using the discourse or document context to improve NMT, in an online setting, by using the past context (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2017; Voita et al., 2018), and in an offline setting, using the past and future context (Maruf and Haffari, 2018). In this paper, we design and evaluate a conversational Bi-MSMT model, where we incorporate the source and target-side conversation histories into a sentence-based attentional model (Bahdanau et al., 2015). Here, the source history comprises of sentences in the original language for both languages, and the target history consists of their corresponding translations. We experiment with different ways of computing the source context representation for this task. Furthermore, we present an effective approach to leverage the target-side context, and also present an intuitive approach for incorporating both contexts simultaneously. To evaluate this task, we introduce datasets extracted from Europarl v7 and OpenSubtitles2016, containing speaker information. Our experiments on English-French, English-Estonian, English-German and English-Russian language-pairs show improvements of +1.44, +1.16, +1.75 and +0.30 BLEU, respectively, for our best model over the context-free baseline. The results show the impact of conversation history on translation of bilingual multi-speaker conversations and can be used as benchmark for future work on this task.

2 Related Work

Our research builds upon prior work in the field of context-based language modelling and context-based machine translation.

Language Modelling There have been few works on leveraging context information for language modelling. Ji et al. (2015) introduced Document Context Language Model (DCLM) which incorporates inter and intra-sentential contexts. Hoang et al. (2016) make use of side information, e.g. metadata, and Tran et al. (2016) use inter-document context to boost the performance

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 101–112
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64011

of RNN language models.

For conversational language modelling, Ji and Bilmes (2004) propose a statistical multi-speaker language model (MSLM) that considers words from other speakers when predicting words from the current one. By taking the inter-speaker dependency into account using a normal trigram context, they report significant reduction in perplexity.

Statistical Machine Translation The few SMT-based attempts to document MT are either restrictive or do not lead to significant improvements upon automatic evaluation. Few of these deal with specific discourse phenomena, such as resolving anaphoric pronouns (Hardmeier and Federico, 2010) or lexical consistency of translations (Garcia et al., 2017). Others are based on a two-pass approach i.e., to improve the translations already obtained by a sentence-level model (Hardmeier et al., 2012; Garcia et al., 2014).

Neural Machine Translation Using context-based neural models for improving online and offline NMT is a popular trend recently. Jean et al. (2017) extend the vanilla attention-based NMT model (Bahdanau et al., 2015) by conditioning the decoder on the previous source sentence via a separate encoder and attention component. Wang et al. (2017) generate a summary of three previous source sentences via a hierarchical RNN, which is then added as an auxiliary input to the decoder. Bawden et al. (2017) explore various ways to exploit context from the previous sentence on the source and target-side by extending the models proposed by Jean et al. (2017); Wang et al. (2017). Apart from being difficult to scale, they report deteriorated BLEU scores when using the target-side context.

Tu et al. (2017) augment the vanilla NMT model with a continuous cache-like memory, along the same lines as the cache-based system for traditional document MT (Gong et al., 2011), which stores hidden representations of recently generated words as translation history. The proposed approach shows significant improvements over all baselines when translating subtitles and comparable performance for news and TED talks. Along similar lines, Kuang et al. (2018) propose dynamic and topic caches to capture contextual information either from recently translated sentences or the entire document to model coherence for NMT. Voita et al. (2018) introduce a context-aware NMT model in which they control and analyse the flow of information from the extended context to the translation model. They show that using the previous sentence as context their model is able to implicitly capture anaphora.

For the offline setting, Maruf and Haffari (2018) incorporate the global source and target document contexts into the base NMT model via memory networks. They report significant improvements using BLEU and METEOR for the contextual model over the baseline. To the best of our knowledge, there has been no work on Multi-Speaker MT or its variation to date.

3 Preliminaries

3.1 Problem Formulation

We are given a dataset that comprises parallel conversations, and each conversation consists of *turns*. Each turn is constituted by sentences spoken by a single speaker, denoted by $\mathbf{x}$ or $\mathbf{y}$, if the sentence is in English or Foreign language, respectively. The goal is to learn a model that is able to leverage the mixed-language conversation history in order to produce high quality translations.

3.2 Data

Standard machine translation datasets are inappropriate for Bi-MSMT task since they are not composed of conversations or the speaker annotations are missing. In this section, we describe how we extract data from raw Europarl v7 (Koehn, 2005) and OpenSubtitles2016[1] (Lison and Tiedemann, 2016) for this task[2].

Europarl The raw Europarl v7 corpus (Koehn, 2005) contains `SPEAKER` and `LANGUAGE` tags where the latter indicates the language the speaker was actually using. The individual files are first split into conversations. The data is tokenised (using scripts by Koehn (2005)), and cleaned (headings and single token sentences removed). Conversations are divided into smaller ones if the number of speakers is greater than 5.[3] The corpus is then randomly split into train/dev/test sets with respect to conversations in ratio 100:2:3. The English side of the corpus is set as reference, and

[1] `http://www.opensubtitles.org/`

[2] The data is publicly available at `https://github.com/sameenmaruf/Bi-MSMT.git`

[3] Using the conversations as is or setting a higher threshold further reduces the data due to inconsistencies in conversation/turn lengths in the source and target side.

	Europarl			Subtitles
	En-Fr	En-Et	En-De	En-Ru
# Conversations	6997	4394	3582	23126
# Sentences	246540	174218	109241	291516
Mean Statistics per Conversation				
# Sentences	36.24	40.65	31.50	13.60
# Turns	4.77	4.85	4.79	7.12
Turn Length	7.12	7.92	6.16	1.68

Table 1: General statistics for training set.

if the language tag is absent, the source language is English, otherwise Foreign. The sentences in the source-side of the corpus are kept or swapped with those in the target-side based on this tag.

We perform the aforementioned steps for English-French, English-Estonian and English-German, and obtain the bilingual multi-speaker corpora for the three language pairs. Before splitting into train/dev/test sets, we remove conversations with sentences having more than 100 tokens for English-French, English-German and more than 80 tokens for English-Estonian[4] respectively, to limit the sentence-length for using subwords with BPE (Sennrich et al., 2016). The data statistics are given in Table 1 and Appendix A[5].

Subtitles There has been recent work to obtain speaker labels via automatic turn segmentation for the OpenSubtitles2016 corpus (Lison and Meena, 2016; van der Wees et al., 2016; Wang et al., 2016). We obtain the English side of OpenSubtitles2016 corpus annotated with speaker information by Lison and Meena (2016).[6] To obtain the parallel corpus, we use the OpenSubtitles alignment links to align foreign subtitles to the annotated English ones. For each subtitle, we extract individual conversations with more than 5 sentences and at least two turns. Conversations with more than 30 turns are discarded. Finally, since subtitles are in a single language, we assign language tag such that the same language occurs in alternating turns. We thus obtain the Bi-MSMT corpus for English-Russian, which is then divided into training, development and test sets.

[4]Sentence-lengths of 100 tokens result in longer sentences than what we get for the other two language-pairs.

[5]Although the extracted dataset is small but we believe it to be a realistic setting for a real-world conversation task, where reference translations are usually not readily available and expensive to obtain.

[6]The majority of sentences still have missing annotations (Lison and Meena, 2016) due to changes between the original script and the actual movie or alignment problems between scripts and subtitles. As for Wang et al. (2016), their publicly released data is even smaller than our En-De dataset extracted from Europarl.

3.3 Sentence-based attentional model

Our base model consists of two sentence-based NMT architectures (Bahdanau et al., 2015), one for each translation direction. Each of them contains an encoder to *read* the source sentence and an attentional decoder to *generate* the target translation one token at a time.

Encoder It maps each source word x_m to a distributed representation $\boldsymbol{h}_m$ which is the concatenation of the corresponding hidden states of two RNNs running in opposite directions over the source sentence. The forward and backward RNNs are taken to be GRUs (gated-recurrent unit; Cho et al. (2014)) in this work.

Decoder The generation of each target word y_n is conditioned on all the previously generated words $\boldsymbol{y}_{<n}$ via the state $\boldsymbol{s}_n$ of the decoder, and the source sentence via a *dynamic* context vector $\boldsymbol{c}_n$:

$$\begin{aligned} y_n &\sim \text{softmax}(\boldsymbol{W}_y \cdot \boldsymbol{u}_n + \boldsymbol{b}_y) \\ \boldsymbol{u}_n &= \tanh(\boldsymbol{s}_n + \boldsymbol{W}_{uc} \cdot \boldsymbol{c}_n + \boldsymbol{W}_{un} \cdot \boldsymbol{E}_T[y_{n-1}]) \\ \boldsymbol{s}_n &= \text{GRU}(\boldsymbol{s}_{n-1}, \boldsymbol{E}_T[y_{n-1}], \boldsymbol{c}_n) \end{aligned}$$

where $\boldsymbol{E}_T[y_{n-1}]$ is the embedding of previous target word y_{n-1}, and $\{\boldsymbol{W}_{(\cdot)}, \boldsymbol{b}_y\}$ are the parameters. The fixed-length *dynamic* context representation of the source sentence $\boldsymbol{c}_n = \sum_m \alpha_{nm} \boldsymbol{h}_m$ is generated by an attention mechanism where $\boldsymbol{\alpha}$ specifies the proportion of relevant information from each word in the source sentence.

4 Conversational Bi-MSMT Model

Before we delve into the details of how to leverage the conversation history, we identify the three types of context we may encounter in an ongoing bilingual multi-speaker conversation, as shown in Figure 1. It comprises of: (i) the previously completed English turns, (ii) the previously completed Foreign turns, and (iii) the ongoing turn (English or Foreign).

We propose a conversational Bi-MSMT model that is able to incorporate all three types of context using source, target or dual conversation histories into the base model. The base model caters to the speaker's language transition by having one sentence-based NMT model (described previously) for each translation direction, English→Foreign and Foreign→English. We now

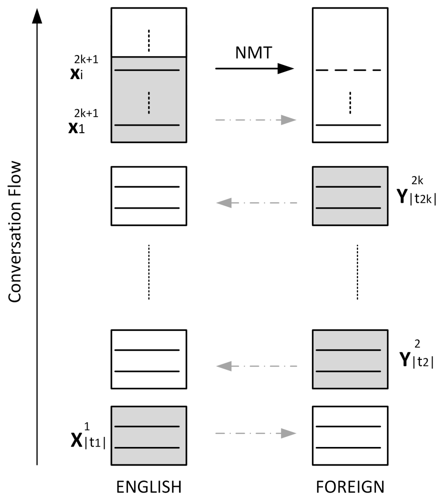

Figure 1: Overview of an ongoing conversation while translating i^{th} sentence in $2k+1^{th}$ turn. $\mathbf{X}^j_{|t_j|}$ and $\mathbf{Y}^j_{|t_j|}$ denote the sentences in previous English and Foreign turn respectively, and $\mathbf{x}^j_i$ denotes the sentence i in ongoing turn j where $i \in \{1, ..., |t_j|\}$. The shaded turns are observed i.e., source (the speaker utterances), while the rest are unobserved i.e., the target translations or the unuttered source sentences for current turn.

describe our approach for extracting relevant information from the source and target bilingual conversation history.

4.1 Source-Side History

Suppose we are translating an ongoing conversation having alternating turns of English and Foreign. We are currently in the $2k+1^{th}$ turn (in English) and want to translate its i^{th} sentence using the source-side conversation history represented by context vector $\mathbf{o}_{src}$ (dimensions H).

Let's assume that we already have the representations of previous source sentences in the conversation. We pass the source sentence representations through Turn-RNNs, which are composed of language-specific bidirectional RNNs irrespective of the speaker, as shown in Figure 2, and concatenate the last hidden states of the forward and backward Turn-RNNs to get the final turn representation $\mathbf{r}_j$, where j denotes the turn index. The individual turn representations are then combined, based on language[7], to obtain context vectors $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$, computed in several possible ways (described below), which are further amalgamated us-

[7]For this work, we define the turns based on language and do not use the speaker information as for real-world chat scenarios (e.g., agent-client in a customer service chat), we do not have multiple speakers based on language. We leave this for future exploration.

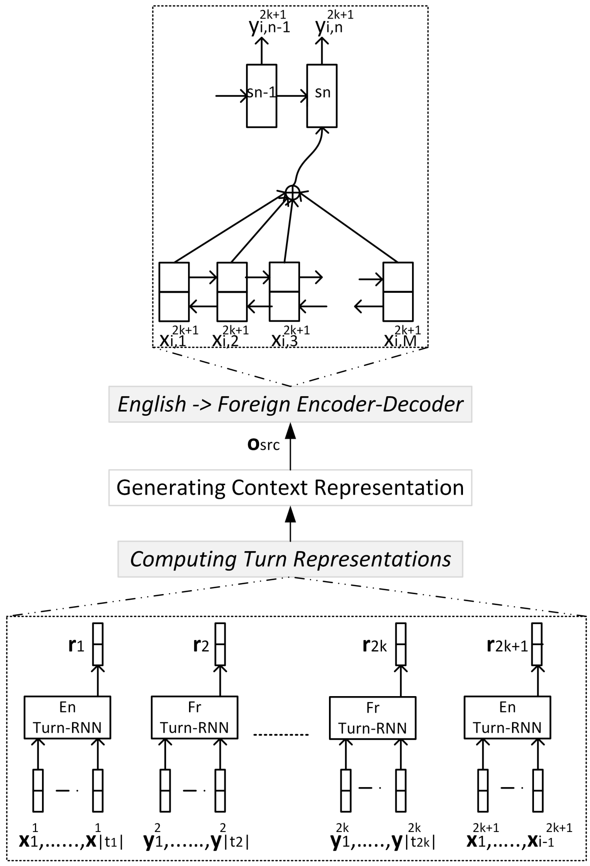

Figure 2: Architectural overview when translating i^{th} sentence in $2k+1^{th}$ turn using source history.

ing a gating mechanism so as to give differing importance to each element of the context vector:

$$\mathbf{o}_{en,fr} = \boldsymbol{\alpha} \odot \mathbf{o}_{en} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{o}_{fr} \quad (1)$$
$$\boldsymbol{\alpha} = \sigma(\mathbf{U}_{en} \times \mathbf{o}_{en} + \mathbf{U}_{fr} \times \mathbf{o}_{fr} + \mathbf{b}_g)$$

where σ is the logistic sigmoid function, $\mathbf{U}$'s are matrices and $\mathbf{b}_g$ is a vector. Finally, we perform a dimensionality reduction to obtain:

$$\mathbf{o}_{src} = \tanh(\mathbf{W}_T \times \mathbf{o}_{en,fr} + \mathbf{b}_T) \quad (2)$$

In the remainder of this section, $\{\mathbf{W}, \mathbf{U}, \mathbf{b}\}$ are language-specific learned parameters. We propose five ways of computing the language-specific context representations, $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$.

Direct Transformation The simplest approach is to combine turn representations using a language-specific dimensionality reduction transformation:

$$\mathbf{o}_{en} = \tanh([\mathbf{W}_{en}; ...; \mathbf{W}_{en}] \times [\mathbf{r}_1; ...; \mathbf{r}_{2k+1}] + \mathbf{b}_{en})$$
$$\mathbf{o}_{fr} = \tanh([\mathbf{W}_{fr}; ...; \mathbf{W}_{fr}] \times [\mathbf{r}_2; ...; \mathbf{r}_{2k}] + \mathbf{b}_{fr})$$

Here $\mathbf{r}_j$'s are concatenated row-wise.

Hierarchical Gating We propose a language-specific exponential decay gating based on the intuition that the farther the previous turns are from the current one, the lesser their impact may be on the translation of a sentence in an ongoing turn, similar in spirit to the caching mechanism by Tu et al. (2017):

$$\mathbf{o}_{en} = g_{en}(g_{en}(...g_{en}(g_{en}(\mathbf{r}_1, \mathbf{r}_3), \mathbf{r}_5)...), \mathbf{r}_{2k-1}), \mathbf{r}_{2k+1})$$

where

$$\begin{aligned} g_{en}(\mathbf{a}, \mathbf{b}) &= \boldsymbol{\alpha} \odot \mathbf{a} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{b} \\ \boldsymbol{\alpha} &= \sigma(\mathbf{U}_{1,en} \times \mathbf{a} + \mathbf{U}_{2,en} \times \mathbf{b} + \mathbf{b}_{en}) \end{aligned}$$

$\mathbf{o}_{fr}$ is computed in a similar way.

Language-Specific Attention The English and Foreign turn representations are combined separately via attention to allow the model to focus on relevant turns in the English and the Foreign context:

$$\begin{aligned} \mathbf{p}_{en} &= \text{softmax}([\mathbf{r}_1; ...; \mathbf{r}_{2k+1}]^T \times \mathbf{h}_i) \qquad (3) \\ \mathbf{p}_{fr} &= \text{softmax}([\mathbf{r}_2; ...; \mathbf{r}_{2k}]^T \times \tanh(\mathbf{W}_{en} \times \mathbf{h}_i + \mathbf{b}_{en})) \\ \mathbf{o}_{en} &= \tanh(\mathbf{W}_{en} \times ([\mathbf{r}_1; ...; \mathbf{r}_{2k+1}] \times \mathbf{p}_{en}) + \mathbf{b}_{en}) \\ \mathbf{o}_{fr} &= [\mathbf{r}_2; ...; \mathbf{r}_{2k}] \times \mathbf{p}_{fr} \end{aligned}$$

Here $\mathbf{r}_j$'s are concatenated column-wise, $\mathbf{h}_i$ is the concatenation of last hidden state of forward and backward RNNs in the encoder for current sentence i in turn $2k+1$ (dimensions *2H*) and $\{\mathbf{W}_{en}, \mathbf{b}_{en}\}$ transform the language space to that of the target language.

Combined Attention This is a language-independent attention that merges all turn representations into one. The hypothesis here is to verify if the model actually benefits from Language-Specific attention or not.

$$\begin{aligned} \mathbf{p}_{en,fr} &= \text{softmax}([\mathbf{r}_{1,en}; \mathbf{r}_2; ...; \mathbf{r}_{2k+1,en}]^T \times \\ &\qquad \tanh(\mathbf{W}_{en} \times \mathbf{h}_i + \mathbf{b}_{en})) \\ \mathbf{o}_{en,fr} &= [\mathbf{r}_{1,en}; \mathbf{r}_2; ...; \mathbf{r}_{2k+1,en}] \times \mathbf{p}_{en,fr} \end{aligned}$$

Here $\mathbf{r}_{2k+1,en} = \tanh(\mathbf{W}_{en} \times \mathbf{r}_{2k+1} + \mathbf{b}_{en})$.

Language-Specific Sentence-level Attention All the previous approaches for computing $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$ use a single turn-level representation. We propose to use the sentence information explicitly via a sentence-level attention to evaluate the significance of more fine-grained context in contrast to Language-Specific Attention. We first concatenate the hidden states of forward and backward Turn-RNNs for each sentence and get a matrix comprising of representations of all the previous source sentences, i.e., for English turns, we have $[\mathbf{r}_1^1; ...; \mathbf{r}_{|t_1|}^1; ...; \mathbf{r}_1^{2k+1}; ...; \mathbf{r}_{i-1}^{2k+1}]$, and similarly we have another matrix for all the previous Foreign sentences. Here, each $\mathbf{r}_i^j$ is the representation of source sentence i in turn j computed by the bidirectional Turn-RNN. The remaining computations are same as in Eq. 3.

4.2 Target-Side History

Using target-side conversation history is as important as that of the source-side since it helps in making the translation more faithful to the target language. This becomes crucial for translating conversations where the previous turns are all in the same language. For incorporating the target-side context, we use a sentence-level attention similar to the one described for the source-side context, i.e., for all previous English source sentences, we have a matrix $\mathbf{R}_{en}$ comprising of the corresponding target sentence representations in Foreign, and another matrix $\mathbf{R}_{fr}$ of target sentence representations (in English) for previous Foreign turns. Here each target sentence representation has dimensions *H*. Then,

$$\begin{aligned} \mathbf{p}_{en} &= \text{softmax}(\mathbf{R}_{en}^T \times \tanh(\mathbf{W}_{t,en} \times \mathbf{h}_i + \mathbf{b}_{t,en})) \\ \mathbf{p}_{fr} &= \text{softmax}(\mathbf{R}_{fr}^T \times (\mathbf{W}_{td,en} \times \mathbf{h}_i + \mathbf{b}_{td,en})) \\ \mathbf{o}_{en} &= \mathbf{R}_{en} \times \mathbf{p}_{en} \\ \mathbf{o}_{fr} &= \tanh(\mathbf{W}_{t,en} \times (\mathbf{R}_{fr} \times \mathbf{p}_{fr}) + \mathbf{b}_{t,en}) \end{aligned}$$

where $\{\mathbf{W}_{t,en}, \mathbf{b}_{t,en}\}$ are for dimensionality reduction and changing the language space of the query vector $\mathbf{h}_i$ and the context vector, while $\{\mathbf{W}_{td,en}, \mathbf{b}_{td,en}\}$ are only for dimensionality reduction. $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$ are further combined using a gating mechanism as in Eq. 1 to obtain the final target context vector $\mathbf{o}_{tgt}$ (dimensions *H*).

4.3 Dual Conversation History

Now that we have explained how to leverage the source and target conversation history separately, we explain how they can be utilised simultaneously. The simplest way to do this is to incorporate both context vectors $\mathbf{o}_{src}$ and $\mathbf{o}_{tgt}$ into the base model (explained in Sec 4.4), referred as *Src-Tgt* dual context.

Another intuitive approach, as evident from Figure 2, is to separately model English and Foreign sentences using two separate context vectors $\mathbf{o}_{en,m}$ and $\mathbf{o}_{fr,m}$, where each is constructed from a mixture of the original source or target translations, is language-specific and possibly contain less noise. We refer to this as the *Src-Tgt-Mix* dual context. Suppose $\mathbf{R}_{en,m}$ contains the mixed source/target representations for English (the dimensions for source representations have been reduced to H) and $\mathbf{R}_{fr,m}$ contains the same for Foreign. Then,

$$\begin{aligned}
\mathbf{p}_{en,m} &= \text{softmax}(\mathbf{R}_{en,m}^T \times (\mathbf{W}_{td,en} \times \mathbf{h}_i + \mathbf{b}_{td,en})) \\
\mathbf{p}_{fr,m} &= \text{softmax}(\mathbf{R}_{fr,m}^T \times \tanh(\mathbf{W}_{tt,en} \times \mathbf{h}_i + \mathbf{b}_{tt,en})) \\
\mathbf{o}_{en,m} &= \tanh(\mathbf{W}_{tr,en} \times (\mathbf{R}_{en,m} \times \mathbf{p}_{en,m}) + \mathbf{b}_{tr,en}) \\
\mathbf{o}_{fr,m} &= \mathbf{R}_{fr,m} \times \mathbf{p}_{fr,m}
\end{aligned}$$

where $\mathbf{W}_{td,en}$, $\mathbf{W}_{tr,en}$ and $\mathbf{W}_{tt,en}$ are for dimensionality reduction, changing the language space and both, respectively.

4.4 Incorporating Context into Base Model

The final representations $\mathbf{o}_{src}$ and $\mathbf{o}_{tgt}$ or $\mathbf{o}_{en,m}$ and $\mathbf{o}_{fr,m}$, can be incorporated together or individually in the base model by:

- **InitDec** Using a non-linear transformation to initialise the decoder, similar to Wang et al. (2017): $\mathbf{s}_{i,0} = \tanh(\mathbf{V} \times \mathbf{o}_i + \mathbf{b}_s)$, where i is the sentence index in current turn $2k+1$, $\{\mathbf{V}, \mathbf{b}_s\}$ are encoder-decoder specific parameters and $\mathbf{o}_i$ is either a single context vector or a concatenation (transformed) of the two.

- **AddDec** As an auxiliary input to the decoder (similar to Jean et al. (2017); Wang et al. (2017); Maruf and Haffari (2018)):

$$\begin{aligned}
\mathbf{s}_{i,n} = \tanh(&\boldsymbol{W}_s \cdot \mathbf{s}_{i,n-1} + \boldsymbol{W}_{sn} \cdot \boldsymbol{E}_T[y_{i,n}] + \\
&\boldsymbol{W}_{sc} \cdot \boldsymbol{c}_{i,n} + \boldsymbol{W}_{ss} \cdot \mathbf{o}_{i,src} + \boldsymbol{W}_{st} \cdot \mathbf{o}_{i,tgt})
\end{aligned}$$

- **InitDec+AddDec** Combination of previous two approaches.

4.5 Training and Decoding

The model parameters are trained end-to-end by maximising the sum of log-likelihood of the bilingual conversations in training set $\mathcal{D}$. For example, for a conversation having alternating turns of English and Foreign language, the log-likelihood is:

$$\sum_{k=0}^{\frac{|T|}{2}-1} \Big(\sum_{i=1}^{|t_{2k+1}|} \log P_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \mathbf{o}_i) + \sum_{j=1}^{|t_{2k+2}|} \log P_{\boldsymbol{\theta}}(\boldsymbol{x}_j|\boldsymbol{y}_j, \mathbf{o}_j) \Big)$$

where i, j denote sentences belonging to $2k+1^{th}$ or $2k+2^{th}$ turn; $\mathbf{o}_{(.)}$ is a representation of the conversation history, and $|T|$ is the total number of turns (assumed to be even here).

The best output sequence for a given input sequence for the i^{th} sentence at test time, a.k.a. decoding, is produced by:

$$\arg\max_{\boldsymbol{y}_i} P_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \mathbf{o}_i)$$

5 Experiments

Implementation and Hyperparameters We implement our conversational Bi-MSMT model in C++ using the DyNet library (Neubig et al., 2017). The base model is built using `mantis` (Cohn et al., 2016) which is an implementation of the generic sentence-level NMT model using DyNet.

The base model has single layer bidirectional GRUs in the encoder and 2-layer GRU in the decoder[8]. The hidden dimensions and word embedding sizes are set to 256, and the alignment dimension (for the attention mechanism in the decoder) is set to 128.

Models and Training We do a stage-wise training for the base model, i.e., we first train the English→Foreign architecture and the Foreign→English architecture, using the sentence-level parallel corpus. Both architectures have the same vocabulary[9] but separate parameters to avoid biasing the embeddings towards the architecture trained last. The contextual model is pre-trained similar to training the base model. The best model is chosen based on minimum overall perplexity on the bilingual dev set.

For the source context representations, we use the sentence representations generated by two sentence-level bidirectional RNNLMs (one each for English and Foreign) trained offline. For the target sentence representations, we use the last hidden states of the decoder generated from the pre-trained base model[10]. At decoding time, however, we use the last hidden state of the decoder computed by our model (not the base) as the target sentence representations. Further training details are provided in Appendix B.

[8]We follow Cohn et al. (2016) and Britz et al. (2017) in choosing hyperparameters for our model.

[9]For each language-pair, we use BPE (Sennrich et al., 2016) to obtain a joint vocabulary of size ≈30k.

[10]Even though the paramaters of the base model are updated, the target sentence representations are fixed throughout training. We experimented with a scheduled updating scheme in preliminary experiments but it did not yield significant improvement.

	Europarl									Subtitles		
	En-Fr			En-Et			En-De			En-Ru		
	Overall	En→Fr	Fr→En	Overall	En→Et	Et→En	Overall	En→De	De→En	Overall	En→Ru	Ru→En
Base Model	37.36	38.13	36.03	20.68	18.64	26.65	24.74	21.80	27.74	19.05	14.90	23.04
+Source Context as Lang-Specific Attention via												
InitDec	38.40†	39.19†	36.86†	**21.79**†	19.54†	**28.33**†	**26.34**†	**23.31**†	29.39†	18.88	14.89	22.56
AddDec	38.50†	**39.35**†	36.98†	21.65†	**19.66**†	27.48†	26.30†	23.09†	**29.52**†	19.34	15.16	23.12
InitDec+AddDec	**38.55**†	39.34†	**37.14**†	21.49†	19.43†	27.55†	26.25†	23.18†	29.30†	**19.35**	**15.16**	**23.14**
+Source Context via												
Direct Tranformation	38.35†	39.13†	36.96†	21.75†	**19.59**†	28.07†	26.29†	23.34†	29.22†	19.09	14.89	22.76
Hierarchical Gating	38.33†	39.14†	36.89†	21.62†	19.55†	27.64†	26.31†	23.17†	29.45†	19.20	15.10	22.73
Lang-Specific Attention	38.40†	39.19†	36.86†	21.79†	19.54†	28.33†	26.34†	23.31†	29.39†	**19.35**	**15.16**	**23.14**
Combined Attention	**38.50**†	**39.36**†	36.94†	21.66†	19.52†	27.90†	26.38†	23.31†	29.44†	18.96	14.82	22.92
Lang-Specific S-Attention	38.46†	39.24†	**37.06**†	**21.84**†	19.58†	**28.43**†	**26.49**†	**23.49**†	**29.49**†	19.09	14.59	22.98
+Lang-Specific S-Attention using												
Source Context	38.46†	39.24†	37.06†	**21.84**†	19.58†	**28.43**†	**26.49**†	**23.49**†	29.49†	19.09	14.59	22.98
Target Context	38.76†	**39.57**†	37.35†	21.77†	**19.68**†	27.86†	26.21†	23.16†	29.26†	19.23	14.77	**23.23**
Dual Context Src-Tgt	**38.80**†	39.51†	**37.50**†	21.74†	19.60†	27.98†	26.39†	23.28†	**29.50**†	18.89	14.52	23.06
Dual Context Src-Tgt-Mix	38.76†	39.52†	37.43†	21.68†	19.63†	27.71†	26.37†	23.26†	29.48†	**19.26**	**14.86**	23.01

Table 2: BLEU scores for the bilingual test sets. Here all contexts are incorporated as InitDec for Europarl and InitDec+AddDec for Subtitles unless otherwise specified. **bold**: Best performance, †: Statistically significantly better than the base model, based on bootstrap resampling (Clark et al., 2011) with $p < 0.05$.

5.1 Results

Firstly, we evaluate the three strategies for incorporating context: InitDec, AddDec, InitDec+AddDec, and report the results for source context using Language-Specific Attention in Table 2. For the Europarl data, we see decent improvements with InitDec for En-Et (+1.11 BLEU) and En-De (+1.60 BLEU), and with InitDec+AddDec for En-Fr (+1.19 BLEU). We also observe that, for all language-pairs, both translation directions benefit from context, showing that our training methodology was indeed effective. On the other hand, for the Subtitles data, we see a maximum improvement of +0.30 BLEU for InitDec+AddDec . We narrow down to three major reasons: (i) the data is noisier when compared to Europarl, (ii) the sentences are short and generic with only 1% having more than 27 tokens, and finally (iii) the turns in OpenSubtitles2016 are short compared to those in Europarl (see Table 1), and we show later (Section 5.2) that the context from current turn is the most important.

The next set of experiments evaluates the five different approaches for computing the source-side context. It is evident from Table 2 that for English-Estonian and English-German, our model indeed benefits from using the fine-grained sentence-level information (Language-Specific Sentence-level Attention) as opposed to just the turn-level one.

Finally, our results with source, target and dual contexts are reported. Interestingly, just using the source context is sufficient for English-Estonian and English-German. For English-French, on the other hand, we see significant improvements for the models using the target-side conversation history over using only the source-side. We attribute this to the base model being more efficient and able to generate better translations for En-Fr as it had been trained on a larger corpus as opposed to the other two language-pairs. Unlike Europarl, for Subtitles, we see improvements for our Src-Tgt-Mix dual context variant over the Src-Tgt one for En→Ru, showing this to be an effective approach when the target representations are noisier.

To summarise, for majority of the cases our Language-Specific Sentence-level Attention is a winner or a close second. Using the Target Context is useful when the base model generates reasonable-quality translations; otherwise, using the Source Context should suffice.

Local Source Context Model Most of the previous works for online context-based NMT consider only a single previous sentence as context (Jean et al., 2017; Bawden et al., 2017; Voita et al., 2018). Drawing inspiration from these works, we evaluate our model (trained with Language-Specific Sentence-Level Attention) on the same

	Europarl			Subtitles
	En-Fr	**En-Et**	**En-De**	**En-Ru**
Prev Sent	38.15	21.70	26.09	**19.13**
Our Model	**38.46**†	**21.84**	**26.49**†	19.09

Table 3: BLEU scores for the bilingual test sets. **bold**: Best performance, †: Statistically significantly better than the contextual baseline.

Type of Context	**BLEU**
No context (Base Model)	24.74
Current Turn	26.39
Current Language from Previous Turns	26.21
Other Language from Previous Turns	26.32
Complete Context	**26.49**

Table 4: BLEU scores for En-De bilingual test set.

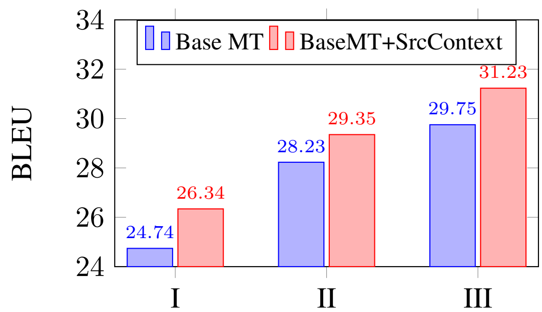

Figure 3: BLEU scores on En-De test set while training (I) smaller base model with smaller corpus (previous experiment), (II) smaller base model with larger corpus, and (III) a larger base model with larger corpus.

test set but using only the previous source sentence as context. This evaluation allows us to hypothesise how much of the gain can be attributed to the previous sentence. From Table 3, it can be seen that our model surpasses the local-context baseline for Europarl showing that the wider context is indeed beneficial if the turn lengths are longer. For En-Ru, it can be seen that using previous sentence is sufficient due to short turns (see Table 1).

5.2 Analysis

Ablation Study We conduct an ablation study to validate our hypothesis of using the complete context versus using only one of the three types of contexts in a bilingual multi-speaker conversation: (i) current turn, (ii) previous turns in current language, and (iii) previous turns in the other language. The results for En-De are reported in Table 4. We see decrease in BLEU for all types of contexts with significant decrease when considering only current language from previous turns.The results show that the current turn has the most influence on translating a sentence, and we conclude

En→Fr	les; par; est; a; dans; le; en; j'; un; afin; question; entre; qu'; être; ces; également; y; depuis; c'; ou
Fr→En	this; of; we; issue; europe; by; up; make; united; does; what; regard; s; must; however; such; whose; share; like; been
En→Et	eest; vahel; üle; nimel; ja; aastal; aasta; neid; ainult seepärast; nagu; kes; komisjoni; tehtud; küsimuses; sisserände; liikmesriigi; mulla; liibanoni; dawit
Et→En	for; this; of; is; political; important; culture; also; as; order; are; each; their; only; gender; were; its; economy; one; market
En→De	daß; auf; und; werden; nicht; müssen; aus; mehr; können; einem; rates; eines; insbesondere; wurden; habe; mitgliedstaaten; ist; sondern; europa; gemeinsamen
De→En	that; its; say; must; some; therefore; more; countries; an; favour; public; will; without; particularly; hankiss; much; increase; eu; them; parliamentary

Table 5: Most frequent tokens correctly generated by our model when compared to the base model.

that since our model is able to capture the complete context, it is generalisable to any conversational scenario.

Training base model with more data To analyse if the context is beneficial even when using more data, we perform an experiment for English-German where we train the base model with additional sentence-pairs from the full WMT'14 corpus[11] (excluding our dev/test sets and filtering sentences with more than 100 tokens). For training the contextual model, we still use the bilingual multi-speaker corpus. We observe a significant improvement of +1.12 for the context-based model (Figure 3 II), showing the significance of conversation history in this experiment condition.[12]

We perform another experiment where we use a larger base model, having almost double the number of parameters than our previous base model (hidden units and word embedding sizes set to 512, and alignment dimension set to 256), to test if the model parameters are being overestimated due to the additional context. We use the same WMT'14 corpus to train the base model and achieve significant improvement of +1.48 BLEU for our context-based model over the larger baseline (Figure 3 III).

[11]https://nlp.stanford.edu/projects/nmt/

[12]It should be noted that the BLEU score for the base model trained with WMT does not match the published results exactly as the test set contains both English and German sentences. It does, however, fall between the scores usually obtained on WMT'14 for En→De and De→En.

Context	nous sommes également favorables au principe d'un système de collecte des miles commun pour le parlement européen, pour que celui-ci puisse bénéficier de billets d'avion moins chers, même si nous voyons difficilement comment ce système pourrait être déployé en pratique. enfin, nous ne sommes pas opposés à l'attribution de prix culturels par le parlement européen.
Source	néanmoins, nous sommes particulièrement critiques à l'égard du prix pour le journalisme du parlement européen et nous ne pensons pas que celui-ci puisse décerner des prix aux journalistes ayant pour mission de soumettre le parlement européen à un regard critique.
Target	however, we are highly critical of parliament's prize for journalism, and do not believe that it is appropriate for parliament to award prizes to journalists whose task it is to critically examine the european parliament.
Base Model	nevertheless, we are particularly critical of the price for the european union's european alism and we do not believe that it would be able to make a price to the journalists who have been made available to the european parliament to a critical view.
Our Model	however, we are particularly critical of the price for the european union's democratic alism and we do not believe that it can give rise to the prices for journalists who have been tabled to submit the european parliament to a critical view.

Table 6: Example En-Fr sentence translation showing how the context helps our model in generating the appropriate discourse connective.

Context	oleks hea, kui reitinguagentuurid vastutaksid tulevikus enda tegevuse eest rohkem. ... kirjalikult. - (it) kiites heaks wolf klinzi raporti, mille eesmärk on reitinguagentuuride tõhus reguleerimine, võtab parlament järjekordse sammu finantsturgude suurema läbipaistvuse suunas. ... mul oli selle dokumendi üle hea meel, sest krediidireitingute valdkonnal on palju probleeme, millest kõige suuremad on oligopolidele tüpilised struktuurid ning konkurentsi, vastutuse ja läbipaistvuse puudumine.
Source	selles suhtes tuleb rõhutada nende tegevuse suuremal äbipaistvuse põhirolli.
Target	in this respect, it is necessary to highlight the central role of increased transparency in their activities.
Base Model	in this regard it must be emphasised in the major role of transparency in which these activities are to be strengthened.
Our Model	in this regard, it must be stressed in the key role of greater transparency in their activities.

Table 7: Example En-Et translation showing how the wide-range context helps in generating the correct pronoun. The antecedent and correct pronoun are highlighted in blue.

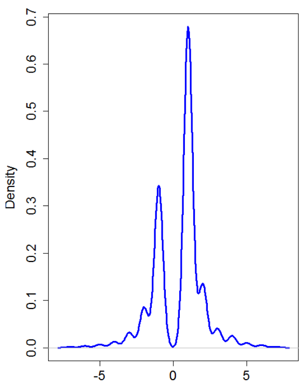

Figure 4: Density of token counts for En$\rightarrow$Fr illustrating where our model is better (+ve x-axis) and where the base model is better (-ve x-axis).

How is the context helping? The underlying hypothesis for this work is that discourse phenomenon in a conversation may depend on long-range dependency and these may be ignored by the sentence-based NMT models. To analyse if our contextual model is able to accurately translate such linguistic phenomenon, we come up with our own evaluation procedure. We aggregate the tokens correctly generated by our model and those correctly generated by the baseline over the entire test set. We then take the difference of these counts and sort them[13]. Table 5 reports the top 20 tokens where our model is better than the baseline for the Europarl dataset. Figure 4 gives the density of counts obtained using our evaluation for En$\rightarrow$Fr[14]. Positive counts correspond to correct translations by our model while the negative counts correspond to where the base model was better. It can be seen that for majority of cases our model supersedes the base model. We observed a similar trend for other translation directions. In general, the correctly generated tokens by our model include pronouns (that, this, its, their, them), discourse connectives (e.g., 'however', 'therefore', 'also') and prepositions (of, for, by).

Table 6 reports an example where our model is able to generate the correct discourse connective '*however*' using the context. If we look at the con-

[13]We do not normalise the counts with the background frequency as it favours rare words. Thus, obscuring the main reasons of improving the BLEU score.

[14]Outliers and tokens with equal counts for our model and the baseline were removed.

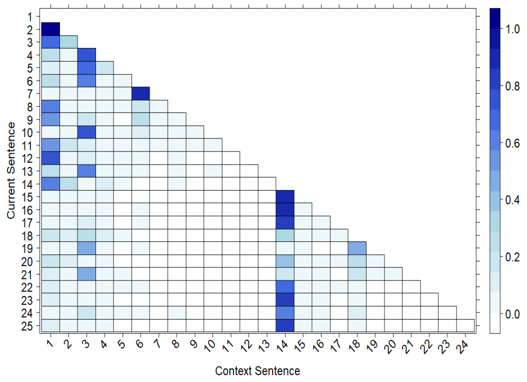

Figure 5: Attention map when translating a conversation from the Et-En test set.

text of the source sentence in French, we come to the conclusion that 'however' is indeed a perfect fit in this case, whereas the base model is at a disadvantage and completely changes the underlying meaning of the sentence by generating the inappropriate connective 'nevertheless'.

Table 7 gives an instance where our model is able to generate the correct pronoun '*their*'. It should be noted that in this case, the current source sentence does not contain the antecedent and thus the context-free baseline is unable to generate the appropriate pronoun. On the other hand, our contextual model is able to do so by giving the highest attention weights to sentences containing the antecedent (observed from the attention map in Figure 5)[15]. Figure 5 also shows that for translating majority of the sentences, the model attends to wide-range context rather than just the previous sentence, hence strengthening the premise of using the complete context.

6 Conclusion

This work investigates the challenges associated with translating multilingual multi-speaker conversations by exploring a simpler task referred to as Bilingual Multi-Speaker Conversation MT. We process Europarl v7 and OpenSubtitles2016 to obtain an introductory dataset for this task. Compared to models developed for similar tasks, our work is different in two aspects: (i) the history captured by our model contains multiple languages, and (ii) our model captures 'global' history as opposed to 'local' history captured in most previous works. Our experiments demonstrate the significance of leveraging the bilingual conversation history in such scenarios. Furthermore, the analysis shows that using wide-range context, our model generates appropriate pronouns and discourse connectives in some cases. We hope this work to be a first step towards translating multilingual multi-speaker conversations. Future work on this task may include optimising the base translation model and approaches that condition on specific discourse information in the conversation history.

[15]For this particular conversation, all previous turns were in Estonian.

Acknowledgments

We thank the reviewers for their useful feedback that helped us in improving our work. We also thank Pierre Lison for providing us with the speaker-annotated OpenSubtitles2016 dataset. This work is supported by the Multi-modal Australian ScienceS Imaging and Visualisation Environment (MASSIVE) (`www.massive.org.au`), the European Research Council (ERC StG DeepSPIN 758969), the Fundação para a Ciência e Tecnologia through contract UID/EEA/50008/2013, a Google Faculty Research Award to GH and the Australian Research Council through DP160102686.

References

Parnia Bahar, Tamer Alkhouli, Jan-Thorsten Peter, Christopher Jan-Steffen Brix, and Hermann Ney. 2017. Empirical investigation of optimization algorithms in neural machine translation. In *Conference of the European Association for Machine Translation*, pages 13–26, Prague, Czech Republic.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL–HLT 2018*.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kyunghyun Cho, B van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Short Papers)*, pages 176–181. Association for Computational Linguistics.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885. Association for Computational Linguistics.

Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108:85–96.

Eva Martínez Garcia, Cristina España-Bonet, and Lluís Màrquez. 2014. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53:103–110.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 909–919. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *International Workshop on Spoken Language Translation*, pages 283–289.

Christian Hardmeier, 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.

Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. 2016. Incorporating side information into recurrent neural network language models. In *Proceedings of NAACL–HLT 2016*, pages 1250–1255.

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

Gang Ji and Jeff Bilmes. 2004. Multi-speaker language modeling. In *Proceedings of HLT–NAACL 2004*.

Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document context language models. In *Workshop track - ICLR 2016*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the 10th Machine Translation Summit*, pages 79–86. AAMT.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. *COLING 2018*.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation of movie & tv subtitles. In *Proceedings of the 2016 Spoken Language Technology Workshop*, pages 245–252, San Diego, CA, USA. IEEE.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Quan Hung Tran, Ingrid Zuckerman, and Gholamreza Haffari. 2016. Inter-document contextual language model. In *Proceedings of NAACL–HLT 2016*, pages 762–766.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2017. Learning to remember translation history with a continuous cache.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2816–2821. Association for Computational Linguistics.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2016. Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of COLING 2016*, pages 2571–2581.

A Data Statistics

	Europarl			Subtitles
	En-Fr	**En-Et**	**En-De**	**En-Ru**
Dev/Test				
# Conversations	140/209	88/132	70/108	462/694
# Sentences	4.9k/7.8k	3.2k/5.2k	2.1k/3.3k	5.9k/9k

Table 8: General statistics for development and test sets.

B Experiments

Training For the base model, we make use of stochastic gradient descent (SGD)[16] with initial learning rate of 0.1 and a decay factor of 0.5 after the fifth epoch for a total of 15 epochs. For the contextual model, we use SGD with an initial learning rate of 0.08 and a decay factor of 0.9 after the first epoch for a total of 30 epochs. To avoid overfitting, we employ dropout and set its rate to 0.2. To reduce the training time of our contextual model, we perform computation of one turn at a time, for instance, when using the source context, we run the Turn-RNNs for previous turns once and re-run the Turn-RNN only for sentences in the current turn.

[16]In our preliminary experiments, we tried SGD, Adam and Adagrad as optimisers, and found SGD to achieve better perplexities in lesser number of epochs (Bahar et al., 2017).