Document-level Neural MT: A Systematic Comparison

António V. Lopes[1] **M. Amin Farajian**[1] **Rachel Bawden**[2]
Michael Zhang[3] **André F. T. Martins**[1]
[1]Unbabel, Rua Visc. de Santarém 67B, Lisbon, Portugal
[2]University of Edinburgh, Scotland, UK
[3] University of Washington, Seattle, WA, USA
{antonio.lopes, amin, andre.martins}@unbabel.com
rachel.bawden@ed.ac.uk, mjqzhang@cs.washington.edu

Abstract

In this paper we provide a systematic comparison of existing and new document-level neural machine translation solutions. As part of this comparison, we introduce and evaluate a document-level variant of the recently proposed Star Transformer architecture. In addition to using the traditional metric BLEU, we report the accuracy of the models in handling anaphoric pronoun translation as well as coherence and cohesion using contrastive test sets. Finally, we report the results of human evaluation in terms of Multidimensional Quality Metrics (MQM) and analyse the correlation of the results obtained by the automatic metrics with human judgments.

1 Introduction

There has been undeniable progress in Machine Translation (MT) in recent years, so much so that for certain languages and domains, when sentences are evaluated in isolation, it has been suggested that MT is on par with human translation (Hassan et al., 2018). However, it has been shown that human translation clearly outperforms MT at the document level, when the whole translation is taken into account (Läubli et al., 2018; Toral et al., 2018; Laubli et al., 2020). For example, the Conference on Machine Translation (WMT) now considers inter-sentential translations in their shared task (Barrault et al., 2019). This sets a demand for context-aware machine translation: systems that take the context into account when translating, as opposed to translating sentences independently.

Translating sentences in context (i.e. at the document level) is essential for correctly handling discourse phenomena whose scope can go beyond the current sentence and which therefore require document context (Hardmeier, 2012; Bawden, 2018; Wang, 2019). Important examples include anaphora, lexical coherence and cohesion, deixis and ellipsis; crucial aspects in delivering high quality translations which often are poorly evaluated using standard automatic metrics.

Numerous context-aware neural MT (NMT) approaches have been proposed in recent years (Tiedemann and Scherrer, 2017; Zhang et al., 2018; Maruf et al., 2019; Miculicich et al., 2018; Voita et al., 2019b; Tu et al., 2018), integrating source-side and sometimes target-side context. However, they have often been evaluated on different languages, datasets, and model sizes. Certain models have also previously been trained on few sentence pairs rather than in more realistic, high-resource scenarios. A direct comparison and analysis of the methods, particularly concerning their individual strengths and weaknesses on different language pairs is therefore currently lacking.

We fill these gaps by comparing a representative set of context-aware NMT solutions under the same experimental settings, providing:

- A systematic comparison of context-aware NMT methods using large datasets (i.e. pre-trained using large amounts of sentence-level data) for three language directions: English (EN) into French (FR), German (DE) and Brazilian Portgueuse (PT_br). We evaluate on (i) document translation using public data for EN→{FR,DE} and (ii) chat translation using proprietary data for all three directions. We use targeted automatic evaluation and human assessments of quality.

- A novel document-level method inspired by the Star transformer approach (Guo et al., 2019), which can leverage full document context from arbitrarily large documents.
- The creation of an additional open-source large-scale contrastive test set for EN→FR anaphoric pronoun translation.[1]

2 Neural Machine Translation

2.1 Sentence-level NMT

NMT systems are based on the encoder-decoder architecture (Bahdanau et al., 2014), where the encoder maps the source sentence into word vectors, and the decoder produces the target sentence given these source representations. These systems, by assuming a conditional independence between sentences, are applied to sentence-level translation, i.e. ignoring source- and target-side context. As such, current state-of-the-art NMT systems optimize the negative log-likelihood of the sentences:

$$p(y^{(k)}|x^{(k)}) = \prod_{t=1}^{n} p(y_t^{(k)}|y_{<t}^{(k)}, x^{(k)}), \quad (1)$$

where $x^{(k)}$ and $y^{(k)}$ are the k^{th} source and target training sentences, and $y_t^{(k)}$ is the t^{th} token in $y^{(k)}$.

In this paper, the underlying architecture is a Transformer (Vaswani et al., 2017). Transformers are usually applied to sentence-level translation, using the sentence independence assumption above. This assumption precludes these systems from learning inter-sentential phenomena. For example, Smith (2017) analyzes certain discourse phenomena that sentence-level MT systems cannot capture, such as obtaining consistency and lexical coherence of named entities, among others.

2.2 Context-aware NMT

Context-aware NMT relaxes the independence assumption of sentence-level NMT; each sentence is translated by conditioning on the current source sentence as well as other sentence pairs (source and target) in the same document. More formally, given a document D containing K sentence pairs $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(K)}, y^{(K)})\}$, the probability of translating $x^{(k)}$ into $y^{(k)}$ is:

$$p(y^{(k)} \mid x^{(k)}) = \prod_{t=1}^{n} p(y_t^{(k)} \mid y_{<t}^{(k)}, X, Y^{(<k)}), \quad (2)$$

where $X := \{x^{(1)}, \ldots, x^{(K)}\}$ are the document's source sentences and $Y^{(<k)} := \{y^{(1)}, \ldots, y^{(k-1)}\}$ the previously generated target sentences.

2.3 Chat translation

A particular case of context-aware MT is chat translation, where the document is composed of utterances from two or more speakers, speaking in their respective languages (Maruf et al., 2018; Bawden et al., 2019).

There are two main defining aspects of chat: the content type (shorter, less planned, more informal and ungrammatical and noisier), and the context available (past utterances only, from multiple speakers in different languages). Specifically, chat is an online task where only the past utterances are available and context-aware models (see §3) need to be adapted to cope with multiple speakers. In this work we introduce tokens to distinguish each speaker and modifying the internal flow of the method to incorporate both speakers' context. There is also an additional challenge in how to handle both language directions and how using gold or predicted context affects chat models. In this work we consider a simplification of this problem by assuming the language direction of the first speaker is always from a gold set, leaving for future work the assessment of the impact of using predictions of the other speaker's utterances.

3 Context-aware NMT methods

We compare three previous context-aware approaches (concatenation, multi-source and cache-based) in our experiments. As well as illustrating different methods of integrating context, they vary in terms of which context (source/target, previous/future) and how much context (number of sentences) they can exploit, as shown in Table 1. Although other context-aware methods do exist, we choose these three methods as being representative of the number of context sentences and usage of both source and target side context.

Concatenation: Tiedemann and Scherrer (2017) use the previous sentence as context, i.e. $X^{(k-1)}$ and $Y^{(k-1)}$, concatenated to the current sentence, i.e. $X^{(k)}$ and $Y^{(k)}$, separated by a special token. It is called `2to1` when just the source-side context is used, and `2to2` when the target is used too.

Multi-source context encoder: Zhang et al. (2018) model the previous source sentences,

[1]The dataset and scripts are available at https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR

$X^{(<k)}$ with an additional encoder. They modify the transformer encoder and decoder blocks to integrate this encoded context; they introduce an additional *context encoder* in the source side that receives the previous two source sentences as context (separated by a special token), encodes them and passes the context encodings to both the encoder and decoder, integrating them using additional multi-head attention mechanisms. Similar to the concatenation-based approach, here the context is limited to the previous few sentences.

Cache-based: Tu et al. (2018) model all previous source and target sentences, $X^{(<k)}$ and $Y^{(<k)}$ with a cache-based approach (Grave et al., 2016), whereby, once a sentence has been decoded, its decoder states and attention vectors are saved in an external key-value memory that can be queried when translating subsequent sentences. This is one of the first approaches that uses the global context.

Other methods have been proposed to use both source and target history with different ranges of context. (Miculicich et al., 2018) attends to words from previous sentences with a 2-stage hierarchical approach, while (Maruf et al., 2019), similarly, attends to words in specific sentences using sparse hierarchical selective attention. (Voita et al., 2019a), which extends the concatenation-based approach to four sentences in a monolingual Automatic Post-Edition (APE) setting; whereas Junczys-Dowmunt (2019) proposes full document concatenation with a BERT model to improve the word embeddings through document context and full document APE. Ng et al. (2019) proposes a noisy channel approach with reranking, where the language model (LM) operates at document-level but the reranking does not. Yu et al. (2019) extends the previous work using conditionally dependent sentence reranking with the document-level LM.

	#Prev	#Fut	Src	Trg
Concat2to1 (1)	1	-	✓	
Concat2to2 (1)	1	-	✓	✓
Multi-source context encoder (2)	2	-	✓	
Cache-based (3)	all	-	✓	✓
Star (4) - (see §4)	all	all (src)	✓	✓
Target APE (5)	3	3		✓
Sparse Hierarchical attn. (6)	all	-	✓	✓

Table 1: A summary of the methods compared (1-4). We also include (5-6) in this summary table for comparative purposes.

4 Doc-Star-Transformer

We propose a scalable approach to document-level NMT inspired by the Star architecture (Guo et al., 2019) for sentence-level NMT. We have an equivalent relay node and build sentence-level representations; we propagate this non-local information at document-level and enrich the word-level embeddings with context information.

To do this, we augment the vanilla sentence-level Transformer model of Vaswani et al. (2017) with two additional multi-headed attention sub-layers. The first sub-layer is used to summarize the global contribution of each sentence into a single embedding. The second layer then uses these sentence embeddings to update word representations throughout the document, thereby incorporating document-wide context.

In §4.1, we describe our model assuming it can attend to context from the entire document without practical memory constraints. Then in §4.2 we show how to extend the model to arbitrarily long contexts by introducing sentence-level recurrence.

4.1 Document-level Context Attention

We begin by describing the encoder of the Doc-Star-Transformer (Figure 1). We refer to the sentence and word representations of the k^{th} sentence at layer i as $\mathbf{s}_i^{(k)}$ and $\mathbf{w}_i^{(k)}$ respectively. Our Doc-Star-Transformer model makes use of the Scaled Dot-Product Attention of Vaswani et al. (2017) to perform alternating updates to sentence and word embeddings across the document to efficiently incorporate document-wide context; our method can efficiently capture local and non-local context (at document-level) and, like the Star Transformer, also eliminates the need to compute pairwise attention scores for each word in the document .

Intermediate word representations, $\mathbf{H}_i^{(k)}$, are updated with sentence-level context. These intermediate representation are then used in a second stage of multi-headed attention to generate an embedding for each sentence in the document.

$$\mathbf{H}_i^{(k)} = \text{Transformer}(\mathbf{w}_{i-1}^{(k)}), \quad (3)$$

$$\mathbf{s}_i^{(k)} = \text{MultiAtt}(\mathbf{s}_{i-1}^{(k)}, \mathbf{H}_i^{(k)}), \quad (4)$$

We then concatenate the newly constructed sentence representations and allow each word in sentence k to attend to all preceding sentences' representations.[2] Finally, we apply a feed-forward net-

[2]We describe our method in the online setting and to match

work, which uses two linear transformations with a ReLU activation to get the layer's final output.

$$\mathbf{H}_{i'}^{(k)} = \text{MultiAtt}(\mathbf{H}_i^{(k)}, [\mathbf{s}_i^{(k)}; \mathbf{s}_i^{(k-1)}; \ldots; \mathbf{s}_i^{(1)}]), \quad (5)$$

$$\mathbf{w}_i^{(k)} = \text{ReLu}(\mathbf{H}_{i'}^{(k)}), \quad (6)$$

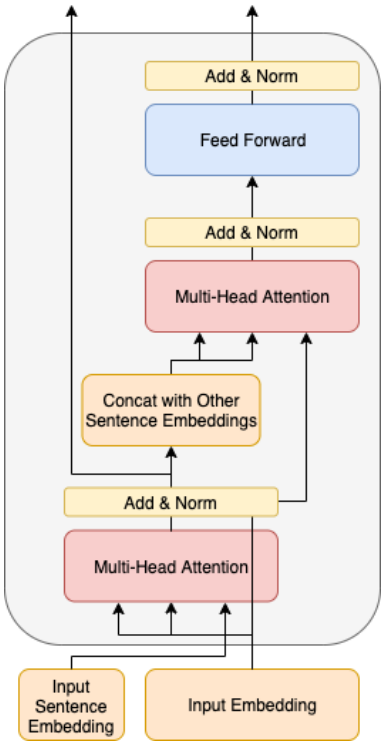

Figure 1: Doc-Star-Transformer encoder.

The Doc-Star-Transformer decoder follows a similar structure to the encoder, except that the decoder does not have access to the sentence representation of the current sentence k, thus, removing sentence $\mathbf{s}_i^{(k)}$ from (5). Source-side context is added through concatenation of the previous sentence embeddings from the final layer of the encoder with the decoder's in (5).

4.2 Sentence-level Recurrence

To overcome practical memory constraints (due to very long documents), we introduce a sentence-level recurrence mechanism with state reuse, similar to that used by Dai et al. (2019). During training, a constant number of sentence embeddings are cached to provide context when translating the next segment. We cut off gradients to these cached sentence embeddings, but allow them to be used to model long-term dependencies without context fragmentation. More formally, we allow τ to be the number of previous sentence embeddings maintained in the cache and update as follows:

$$\mathbf{H}_i^{(k')} = \text{MultiAtt}(\mathbf{H}_i^{(k)}, [\mathbf{s}_i^{(k)}; \mathbf{s}_{i-1}^{(k)}; \ldots; \mathbf{s}_i^{(B)}; \text{SG}(\mathbf{s}_i^{(B)}); \ldots; \text{SG}(\mathbf{s}_i^{(B-\tau)})]),$$

where B is the index of the first sentence in the batch and SGs are the sentence representations with stopped gradients. In contrast with previous approaches, such as Hierarchical Attention (Maruf et al., 2019), this gradient caching strategy has the advantage of letting the model attend to full source context regardless of document lengths and therefore to avoid practical memory issues.

5 Evaluating Context-Aware NMT

The evaluation of context-aware MT is notoriously tricky (Hardmeier, 2012); standard automatic metrics such as BLEU (Papineni et al., 2002) are poorly suited to evaluating discourse phenomena (e.g. anaphoric references, lexical cohesion, deixis, ellipsis) that require document context. We therefore evaluate all models using a range of phenomenon-specific contrastive test sets.

Contrastive sets are an automatic way of evaluating the handling of particular phenomena (Sennrich, 2017; Rios Gonzales et al., 2017). The aim is to assess how well models rank correct translations higher than incorrect (contrastive) ones. For context-aware test sets, the correctness of translations depends on context. Several such sets exist for a range of discourse phenomena and for several language directions: EN→FR (Bawden et al., 2018), EN→DE (Müller et al., 2018) and EN→RU (Voita et al., 2019b). In this article, we evaluate using the following test sets for our two language directions of focus, EN→DE and EN→FR:

EN-FR: anaphora, lexical choice (Bawden et al., 2018):[3] two manually crafted sets (200 contrastive pairs each), for which the previous sentence determines the correct translation. The sets are balanced such that each correct translation also appears as an incorrect one (a non-contextual baseline achieves 50% precision). Anaphora examples include singular and plural personal and possessive pronouns. In addition to standard contrastive examples, this set also contains contextually correct examples, where the antecedent is translated

the decoder side. In the document-MT setting, (5) concatenates all sentences' representations to include context from future source-side sentences during translation.

[3] https://github.com/rbawden/discourse-mt-test-sets

strangely, designed to test the use of past translation decisions. Lexical choice examples include cases of lexical ambiguity (cohesion) and lexical repetition (cohesion).

EN→DE: anaphoric pronouns (ContraPro) (Müller et al., 2018).[4] A large-scale automatically created set from OpenSubtitles2018 (Lison et al., 2018), in which sentences containing the English anaphoric pronoun *it* (and its corresponding German translations *er*, *sie* or *es*) are automatically identified, and contrastive erroneous translations are automatically created. The test set contains 4,000 examples for each target pronoun type, and the disambiguating context can be found in any number of previous sentences.

EN→FR: large-scale pronoun test set We automatically create a large-scale EN→FR test set from OpenSubtitles2018 (Lison et al., 2018) in the style of ContraPro, with some modifications to their protocol due to the limited quality of available tools. The test set is created as follows:

1. Instances of *it* and *they* and their antecedents are detected using NEURALCOREF.[5] Unlike Müller et al. (2018), we only run English coreference due to a lack of an adequate French tool.
2. We align pronouns to their translations (*il*, *elle*, *ils*, *elles*) using FastAlign (Dyer et al., 2013).
3. Examples are filtered to only include subject pronouns (using Spacy[6]) with a nominal antecedent, aligned to a nominal French antecedent matching the pronoun's gender. We also remove examples whose antecedent is more than five sentences away to avoid cases of imprecise coreference resolution.
4. Contrastive translations are created by inverting the pronouns' gender (cf. Figure 2). We modify the gender of words that agree with the pronoun (e.g. adjectives and some past participles) using the Le*fff* lexicon (Sagot, 2010)).

The test set consists of 3,500 examples for each target pronoun type (cf. Table 2 for the distribution of coreference distances).

6 Experimental Setup

As mentioned in §1, we aim to provide a systematic comparison of the approaches over the same datasets, training data sizes and language pairs. We study whether pre-training with larger resources (in a more realistic high-resource scenario) has an impact on the methods on language directions that are challenging for sentence-level MT. We consider translation from English into French (FR), German (DE) and Brazilian Portuguese (PT_br), which all have gendered pronouns corresponding to neuter anaphoric pronouns in English (*it* for all three and *they* for FR and PT_br).

We compare the three previous methods (§3) plus the Doc-Star-Transformer in two scenarios: (i) document MT, testing on TED talks (EN→FR and EN→PT_br), and (ii) chat MT testing on proprietary conversation data for all three directions.

Context sentence	
	Some red **roses** for Your Ladyship.
	Des **roses**$_{\text{fem.}}$ pour madame.
Current sentence	
	Who could **they** be from?
✓	De qui peuvent-**elles**$_{\text{fem.}}$ bien être ?
×	De qui peuvent-**ils**$_{\text{masc.}}$ bien être ?

Figure 2: An example from the large-scale EN→FR test set.

	# examples at each distance					
Pronoun	0	1	2	3	4	5
il	1,628	1,094	363	213	127	75
elle	1,658	1,144	356	166	106	70
ils	1,165	1,180	501	302	196	156
elles	1,535	1,148	409	199	128	81

Table 2: The distribution of each pronoun type according to distance (in #sentences) from the antecedent.

6.1 Data

For both scenarios, we pre-train baseline models on large amounts of publicly available sentence-level parallel data ($\sim 18M$, $\sim 22M$ and $\sim 5M$ sentence pairs for EN→DE, EN→FR, and EN→PT_br respectively). We then separately fine-tune them to each domain. For the document MT task, we consider EN→DE and EN→FR and fine-tune on IWSLT17 (Cettolo et al., 2012) TED Talks, using the test sets 2011-2014 as dev sets, and 2015 as test sets. For the chat MT task, we fine-tune on (anonymized) proprietary data of 3 different domains and on an additional language pair (EN→PT_br). Dataset sizes are shown in Table 3 (sentence-level pre-training data) and Tables 4–5 (document and chat task data respectively).

[4] https://github.com/ZurichNLP/ContraPro
[5] https://github.com/huggingface/neuralcoref
[6] https://spacy.io

	Train	Dev
EN-DE	18M	1K
EN-FR	20M	1K
EN-PT_br	5M	1K

Table 3: Sentence-level corpus sizes (#sentences)

	Train	Dev	Test
EN-DE	206K	5.4K	1.1K
EN-FR	233K	5.8k	1.2K

Table 4: TED talks document-level corpus sizes (#sentences)

		Domain1	Domain2	Domain3
EN-DE	Train	674k	62K	13K
	Dev	37K	3.2K	0.6K
	Test	35K	3.6K	0.7K
EN-FR	Train	395K	108K	110K
	Dev	21K	6.3K	6.1K
	Test	22K	6.2K	6.3K
EN-PT_br	Train	235K	61K	13K
	Dev	13K	3.4K	0.7K
	Test	13K	3.2K	0.7K

Table 5: The corpora sizes of the chat translation task. We consider both speakers for this count.

6.2 Training Configuration

For all experiments we use the *Transformer base* configuration (hidden size of 512, feedforward size of 2048, 6 layers, 8 attention heads) with the learning rate schedule described in (Vaswani et al., 2017). We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017) and early stopping of 5 consecutive non-improving validation points of both accuracy and perplexity. Self-attentive models are sensitive to batch size (Popel and Bojar, 2018), and so we use batches of $32k$ tokens for all methods.[7] For all tasks, we use a subword unit vocabulary (Sennrich et al., 2016) with $32k$ operations. We share source and target embeddings, as well as target embeddings with the final vocab projection layer (Press and Wolf, 2017).

For the document translation experiments, we run the same experimental setting with 3 different seeds and average the scores of each model.

For the approaches that fine-tune just the document-level parameters (i.e. cache-based, multi-source encoder, and Doc-Star-Transformer), we reset all optimizer states and train with the same configuration as the baselines (with the base parameters frozen), as described in (Tu et al., 2018; Zhang et al., 2018). For Doc-Star-Transformer we use multi-heads of 2 and 8 heads. All methods are implemented in Open-NMT (Klein et al., 2017).

[7]The optimizer update is delayed to simulate the $32k$ tokens.

6.3 Chat-specific modifications

In the case of the concatenation-based approaches, multi-source context encoder, and the Doc-Star-Transformer, we add the speaker symbol as special token to the beginning of each sentence. For the cache-based systems, we introduce two different caches, one per speaker, and investigate different methods for deep fusing them (Tu et al., 2018): (i) deep fusing the first speaker's cache first and next fusing with the second speaker's cache, (ii) the same method but with the second speaker first, and (iii) jointly integrating the caches. In addition, for the cache-based system we explore the effect of storing full words or subword units in the external memory For the full word approach, we use subword units in the vocab but merge the words when adding to the cache.

6.4 Evaluation setup

We perform both automatic and manual evaluation, in order to gain more insights into the differences between the models.

Automatic evaluation: We first evaluate all methods with case-sensitive detokenized BLEU (Papineni et al., 2002).[8] We then evaluate context-dependent discourse-level phenomena using the previously described contrastive test sets. For EN→DE this corresponds to the large-scale anaphoric pronoun test set of Müller et al. (2018) and for EN→FR our own analogous large-scale anaphoric pronoun test set (described in §5),[9] as well as the manually crafted test sets of Bawden et al. (2018) for anaphora and coherence/cohesion.

Manual evaluation: In the case of the chat translation task (using proprietary data), in addition to BLEU, we also manually assess the performance of the systems with professional human annotators, who mark the errors of the systems with different levels of severity (i.e. minor, major, critical). In the case of extra-sentential errors such as agreements we asked them to mark both the pronoun and its antecedent. We score the systems' performance using Multidimensional Quality Metrics (MQM) (Lommel, 2013):

$$MQM = 100 - \frac{\text{minor} + \text{major} * 5 + \text{critical} * 10}{\text{Word count}}$$

[8]Using Moses' (Koehn et al., 2007) `multi-bleu-detok`.

[9]For both large-scale test sets, we make sure to exclude the documents they include from the training data.

By having access to the full conversation, the annotators can annotate both intra- and extra-sentential errors (e.g. document-level error examples of agreement or lexical consistency).

We prioritize documents with a large number of edits compared to the sentence-level baseline (normalized by document length) due to document-level systems tending to perform few edits with respect to the high performance non-context-aware systems. We request annotations of approximately 200 sentences per language pair and method.

7 Results and analysis

7.1 Document Translation Task

Table 6 shows the results of the average performance of each system on IWSLT data according to BLEU. Although the approaches have previously shown improved performance compared to a baseline, when a stronger baseline is used, we see marginal to no improvements over the baseline for both language directions.

	EN→DE	EN→FR
Baseline	32.08	40.92
Concat2to1	31.84	40.67
Concat2to2	30.89	40.57
Cache SubWords	32.10	40.91
Cache Words	**32.12**	40.88
Zhang et al. 2018	31.03	**40.95**
Star, 2 heads, gold target ctx	31.76	**41.00**
Star, 2 heads, predicted target ctx	31.39	40.72
Star, 8 heads, gold target ctx	31.74	40.74
Star, 8 heads, predicted target ctx	31.29	40.58

Table 6: BLEU score results on the IWSLT15 test set (averaged over 3 different runs for each method).

Table 7 shows the average performance of each system for all contrastive sets. The results differ greatly from BLEU results; methods on par or below the baseline according to BLEU perform better than the baseline when evaluated on the contrastive test sets. This is notably the case of the Concat models, which achieve some of the best results on the both large-scale pronoun sets (EN→DE and EN→FR), as shown by the high percentages on the more difficult feminine pronoun *Sie* for EN→DE and all pronouns for EN→FR.

Most models struggle to achieve high performances for the feminine *sie* and masculine *er*, which is likely due to neuter *es* being the majority class in the training data. For French, although the feminine pronouns are also usually challenging, the high scores seen here are possibly due to the fact that many examples have an antecedent within the same sentence. The Concat2to2 method however performs well across the board, proving to be an effective way of exploiting context. It also achieves the highest scores on both the anaphora and coherence/cohesion test set, which is only possible when the context is actually being used, as the test set is completely balanced. This appears to confirm the findings of Bawden et al. (2018) that target-side context is most effectively used when channelled through the decoder. Surprisingly, the multi-source encoder approach degrades the baseline with respect to this evaluation, suggesting that the context being used is detrimental to the handling of these phenomena.

We note that using OpenSubtitles as a resource for context-dependent translation or scoring, has additional challenges. Figure 3 illustrates four of these, which could make translation more challenging if they affect the context being exploited.

7.2 Chat Translation Task

Table 8 shows BLEU score results on the proprietary data, with the modifications described in §3 to address the chat task. As expected, document-level information has a larger impact for the lowest resource language pair, EN→PT_br, with marginal improvements on EN→FR and EN→DE.

The performance of these methods depends on the language pair and domain. Although it is not conclusive which method performs best, our proposed method improves over the baseline consistently, whereas the cache-based and Concat2to2 methods also perform well in some scenarios. For our Doc-Star-Transformer approach, using predictions rather than the gold history harms the model at inference, showing that bridging this gap could lead to a better handling of target-side context.

There is little correlation between BLEU scores and the human MQM scores (as shown by the comparison for 3 methods in Table 9). Although the difference between BLEU scores are marginal, MQM indicates that quality differences can be seen by human evaluators: the document-level systems (Cache and Star) both achieve higher results for EN→PT_br (although the Star approach underperforms for EN→FR). This shows that for certain language directions, the document-level approaches do learn to fix some errors and therefore improve translation quality. This also confirms previous suggestions that BLEU is not a good met-

	EN→DE				EN→FR						
	Total	Es	Sie	Er	Total	it		they		Anaphora	Coherence/ cohesion(%)
						elle	il	elles	ils	All	All
Baseline	45.0	91.9	22.9	20.2	79.7	88.1	82.7	76.1	72.2	50.0	50.0
Concat2to1	48.0	91.6	**27.1**	**25.3**	**80.9**	**88.4**	**83.3**	**77.2**	**73.9**	50.0	**52.5**
Concat2to2	70.8	91.8	**61.9**	**58.7**	**83.2**	**89.2**	**86.2**	**80.4**	**77.6**	**82.5**	**55.0**
Cache (Subwords)	45.2	**92.1**	**23.5**	19.9	79.7	88.0	82.7	76.0	72.0	50.0	50.0
Multi-src Enc	42.6	62.3	**33.9**	**31.5**	59.0	62.0	61.3	57.2	57.3	47.0	46.5
Star, 8 heads	45.9	91.3	**27.0**	19.5	79.6	88.0	82.6	76.1	72.0	50.0	50.0

Table 7: Accuracies (in %) for the contrastive sets. Methods outperforming the baseline are in bold.

		Domain1			Domain2			Domain3		
		EN-DE	EN-FR	EN-PT_br	EN-DE	EN-FR	EN-PT_br	EN-DE	EN-FR	EN-PT_br
Baseline		78.53	79.71	81.21	72.11	76	73.94	69.67	74.76	74.95
Concat2to1	S1,S2 + speaker tag	78.04	79.65	80.36	71	75.35	73.02	**69.92**	74.57	74.82
	S1	77.97	79.55	80.26	70.95	75.21	73.33	**69.77**	74.47	74.84
Concat2to2	S1,S2 + speaker tag	**79.84**	79.3	80.33	70.56	74.87	73.52	**69.74**	74.37	74.56
	S1	**78.88**	79.15	79.92	70.13	74.9	73.33	69.59	74.25	74.33
Cache S1 + Cli	JointPolicy Subwords	**78.62**	79.66	80.79	**72.12**	75.03	73.47	69.47	74.77	75.04
	JointPolicy Words	78.52	79.63	80.93	71.66	75.93	73.54	69.55	74.77	74.97
Cache S1 only	Subwods	78.41	79.46	81.17	71.73	75.92	**74.41**	69.68	**74.8**	74.94
	Words	78.28	79.54	81.04	71.9	75.87	**74.33**	69.51	**74.82**	74.94
Multi-src enc	SEP + speaker tag	78.23	79.64	81.04	71.5	75.87	73.78	-	74.66	74.82
Star	S1,S2 2 heads Gold target ctx	**79.7**	**80.08**	**82.64**	71.79	75.62	73.67	**71.36**	**74.87**	**75.03**
	S1,S2 2 heads Predicted target ctx	**78.81**	79.38	79.63	71.72	75.58	73.7	**69.38**	74.77	**75.11**
	S1 2 heads Gold target ctx	**79.35**	79.58	**82.52**	72.16	75.95	**74.1**	**71.33**	**75.01**	**75.48**
	S1 2 heads Predicted target ctx	78.17	79.24	79.83	**72.24**	75.68	73.9	**70.24**	74.65	**75.21**

Table 8: BLEU scores on the chat translation task (proprietary data for 3 different domains and language pairs). S1 and S2 refer to the speakers in the case of chat translation task.

	EN→FR		EN→PT_br	
	BLEU	MQM	BLEU	MQM
Baseline	74.76	87.46	74.95	92.47
Cache	74.82	**89.02**	74,94	**93.20**
Star 2 heads	75.01	86.80	75.48	**95.20**

Table 9: The results of automatic and manual evaluation of the context-aware NMT methods in terms of BLEU and MQM on English→French and English→Portuguese.

ric to distinguish between strong NMT systems.

8 Conclusion

We provided a systematic comparison of several context-aware NMT methods. One of the methods in this comparison was a new adaptation of the recently proposed StarTransformer architecture to document-level MT. In addition to BLEU, we reported results of the contrastive evaluation of context-dependent phenomena (anaphora and coherence/cohesion), creating an additional large-scale contrastive test set for EN→FR anaphoric pronouns, and we carried out human evaluation in terms of Multidimensional Quality Metrics (MQM). Our findings suggest that existing context-aware approaches are less advantageous in scenarios with larger datasets and strong sentence-level baselines. In terms of the targeted context-dependent evaluation, one of the promising approaches is one of the simplest: the Concat2to2, where translated context is channelled through the decoder, although our Doc-Star-Transformer method achieves good results according to the manual evaluation of MT quality.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work is supported by the EU in the context of the PT2020 project (contracts 027767 and 038510) and the H2020 GoURMET project (825299), by the European Research Council (ERC StG DeepSPIN 758969), by the Fundação para a Ciência e Tecnologia through contract UID/EEA/50008/2019 and by the UK Engineering and Physical Sciences Research Council (MTStretch fellowship grant EP/S001271/1).

References

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Difficulty	English	French
Colloquialisms	Well, they just ain't a-treatin' me right	Eh bien, elles me traitent mal 'Well, they're treating me badly'
Paraphrasing	Do not forget your friends, they are always with you heart and soul!	N'oubliez pas vos amis: ils sont toujours près de vous! 'Don't forget your friends: they are always near to you'
Truncation	Neighbor. what have you done?	Voisin ? 'Neighbour?'
Free translation	I don't understand either.	Moi non plus. 'me neither'

Figure 3: Examples of four challenges for MT of OpenSubtitles: (i) colloquialisms, (ii) paraphrasing, (iii) subtitle truncation (can be due to space constraints), and (iv) free translations that fulfill the same discursive role.

Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the 4th Conference on Machine Translation*.

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies*.

Bawden, Rachel, Sophie Rosset, Thomas Lavergne, and Eric Bilinski. 2019. DiaBLa: A Corpus of Bilingual Spontaneous Written Dialogues for Machine Translation.

Bawden, Rachel. 2018. *Going beyond the sentence: Contextual Machine Translation of Dialogue*. Ph.D. thesis, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In *Proceedings of the 57th annual meeting on association for computational linguistics*.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Grave, Edouard, Armand Joulin, and Nicolas Usunier. 2016. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*.

Guo, Qipeng, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hardmeier, Christian. 2012. Discourse in Statistical Machine Translation. a survey and a case study. *Discours*, 11.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*.

Junczys-Dowmunt, Marcin. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Laubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Lison, Pierre, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Lommel, Arle Richard. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. In *Proc. of the 3rd Conference on Machine Translation*.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Müller, Mathias, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation*.

Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.

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*.

Pereyra, Gabriel, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the 5th International Conference on Learning Representations*.

Popel, Martin and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1).

Press, Ofir and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Rios Gonzales, Annette, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the 2nd Conference on Machine Translation*.

Sagot, Benoît. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

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

Sennrich, Rico. 2017. How Grammatical is Character-level Neural Machine Translation? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Smith, Karin Sim. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Tiedemann, Jörg and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*.

Tu, Zhaopeng, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6.

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*.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Wang, Longyue. 2019. *Discourse-Aware Neural Machine Translation*. Ph.D. thesis, Dublin City University, Dublin, Ireland.

Yu, Lei, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2019. Putting machine translation in context with the noisy channel model.

Zhang, Jiacheng, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.