Document-level Neural MT: A Systematic Comparison

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Document-level Neural MT: A Systematic Comparison

António V. Lopes1  M. Amin Farajian1  Rachel Bawden2
Michael Zhang3  André F. T. Martins1
1Unbabel, Rua Visc. de Santarém 67B, Lisbon, Portugal
2University of Edinburgh, Scotland, UK
3University 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 (Hasan 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 Portuguese (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.

© 2020 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.
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:

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

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:

\[ p(y^{(k)} | x^{(k)}) = \prod_{t=1}^{n} p(y_t^{(k)} | y_{<t}^{(k)}, X, Y^{(<k)}) , \]

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 \( 2\to1 \) when just the source-side context is used, and \( 2\to2 \) when the target is used too.

Multi-source context encoder: Zhang et al. (2018) model the previous source sentences,
\(X(<k)\) 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-Editing (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.

### 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 \(s_i(k)\) and \(w_i(k)\) 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, \(H_i(k)\), 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.

\[
H_i(k) = \text{Transformer}(w_i(k)),
\]

\[
s_i(k) = \text{MultiAtt}(s_{i-1}(k), H_i(k)),
\]

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-

---

<table>
<thead>
<tr>
<th>Method</th>
<th>#Prev</th>
<th>#Fut</th>
<th>Src</th>
<th>Trg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concat2to1 (1)</td>
<td>1</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Concat2to2 (1)</td>
<td>1</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multi-source context encoder (2)</td>
<td>2</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cache-based (3)</td>
<td>all</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Star (4) - (see §4)</td>
<td>all</td>
<td>all (src)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Target APE (5)</td>
<td>3</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sparse Hierarchical attn. (6)</td>
<td>all</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

\(^2\) 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.

\[ H_i^{(k)} = \text{MultiAtt}(H_i^{(k)}, [s_i^{(k)}; s_i^{(k-1)}; \ldots; s_i^{(1)}]) \]

\[ w_i^{(k)} = \text{ReLU}(H_i^{(k)}) \]

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 \( s_i^{(k)} \) 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 \( \tau \) to be the number of previous sentence embeddings maintained in the cache and update as follows:

\[ H_i^{(k')} = \text{MultiAtt}(H_i^{(k)}, [s_i^{(k)}; s_i^{(k-1)}; \ldots; s_i^{(B)}; \text{SG}(s_i^{(B)}); \ldots; \text{SG}(s_i^{(B-\tau)}))] \]

where \( B \) is the index of the first sentence in the batch and \( \text{SGs} \) 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 (Hartmeier, 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

3https://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). 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. 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) 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 Lefff 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.

#### 6.1 Data

For both scenarios, we pre-train baseline models on large amounts of publicly available sentence-level parallel data (~18M, ~22M and ~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).
### Table 3: Sentence-level corpus sizes (#sentences)

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-DE</td>
<td>18M</td>
<td>1K</td>
<td></td>
</tr>
<tr>
<td>EN-FR</td>
<td>20M</td>
<td>1K</td>
<td></td>
</tr>
<tr>
<td>EN-PT_br</td>
<td>5M</td>
<td>1K</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Domain1</th>
<th>Domain2</th>
<th>Domain3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-DE</td>
<td>Train</td>
<td>674K</td>
</tr>
<tr>
<td></td>
<td>Dev</td>
<td>37K</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>35K</td>
</tr>
<tr>
<td>EN-FR</td>
<td>Train</td>
<td>395K</td>
</tr>
<tr>
<td></td>
<td>Dev</td>
<td>21K</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>22K</td>
</tr>
<tr>
<td>EN-PT_br</td>
<td>Train</td>
<td>235K</td>
</tr>
<tr>
<td></td>
<td>Dev</td>
<td>13K</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>13K</td>
</tr>
</tbody>
</table>

### 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. 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).

### 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). 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), 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):

\[
MQM = 100 - \frac{\text{Word count}}{-\text{minor} + \text{major} \times 5 + \text{critical} \times 10}
\]

---

8Using Moses’ (Koehn et al., 2007) multi-bleu-detok.

9For 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.

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

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

<table>
<thead>
<tr>
<th></th>
<th>Domain1</th>
<th>Domain2</th>
<th>Domain3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN→DE</td>
<td>Domains</td>
<td>EN→FR</td>
</tr>
<tr>
<td>Baseline</td>
<td>78.53</td>
<td>79.71</td>
<td>81.21</td>
</tr>
<tr>
<td>Concat2to1</td>
<td>78.04</td>
<td>79.63</td>
<td>80.36</td>
</tr>
<tr>
<td>Concat2to2</td>
<td>78.88</td>
<td>79.13</td>
<td>79.92</td>
</tr>
<tr>
<td>Cache S1 + Subwords</td>
<td>78.62</td>
<td>79.66</td>
<td>80.79</td>
</tr>
<tr>
<td>Cache S1 + JointPolicy</td>
<td>78.52</td>
<td>79.63</td>
<td>80.93</td>
</tr>
<tr>
<td>Cache S1</td>
<td>78.41</td>
<td>79.46</td>
<td>81.17</td>
</tr>
<tr>
<td>Multi-src Enc</td>
<td>78.28</td>
<td>79.54</td>
<td>81.04</td>
</tr>
<tr>
<td>Star with 2 heads</td>
<td>78.23</td>
<td>79.64</td>
<td>81.04</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>EN→FR</th>
<th>EN→PT_Jr</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU MQM BLEU MQM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>74.76</td>
<td>87.46</td>
</tr>
<tr>
<td>Cache</td>
<td>74.82</td>
<td>89.02</td>
</tr>
<tr>
<td>Star 2 heads</td>
<td>75.01</td>
<td>86.80</td>
</tr>
</tbody>
</table>

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.

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, 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


<table>
<thead>
<tr>
<th>Difficulty</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquialisms</td>
<td>Well, they just ain’t a-treatin’ me right</td>
<td>Eh bien, elles me traitent mal</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>Do not forget your friends, they are always with you heart and soul!</td>
<td>N’oubliez pas vos amis: ils sont toujours près de vous!</td>
</tr>
<tr>
<td>Truncation</td>
<td>Neighbor. what have you done?</td>
<td>Vosin ?</td>
</tr>
<tr>
<td>Free translation</td>
<td>I don’t understand either.</td>
<td>Mon non plus.</td>
</tr>
</tbody>
</table>

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.


