A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation

Mathias Müller[1,2] **Annette Rios**[1] **Elena Voita**[3,4] **Rico Sennrich**[1,5]

[1]Institute of Computational Linguistics, University of Zurich
[2]Amazon Research, Berlin*
[3]Yandex Research, Russia [4]University of Amsterdam, Netherlands
[5]School of Informatics, University of Edinburgh

Abstract

The translation of pronouns presents a special challenge to machine translation to this day, since it often requires context outside the current sentence. Recent work on models that have access to information across sentence boundaries has seen only moderate improvements in terms of automatic evaluation metrics such as BLEU. However, metrics that quantify the overall translation quality are ill-equipped to measure gains from additional context. We argue that a different kind of evaluation is needed to assess how well models translate inter-sentential phenomena such as pronouns. This paper therefore presents a test suite of contrastive translations focused specifically on the translation of pronouns. Furthermore, we perform experiments with several context-aware models. We show that, while gains in BLEU are moderate for those systems, they outperform baselines by a large margin in terms of accuracy on our contrastive test set. Our experiments also show the effectiveness of parameter tying for multi-encoder architectures.

1 Introduction

Even though machine translation has improved considerably with the advent of neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015), the translation of pronouns remains a major issue. They are notoriously hard to translate since they often require context outside the current sentence.

As an example, consider the sentences in Figure 1. In both languages, there is a pronoun in the second sentence that refers to the European Central Bank. When the second sentence is translated from English to German, the translation of the pronoun *it* is ambiguous. This ambiguity can only be resolved with context awareness: if a translation system has access to the previous English sentence, the previous German translation, or both, it can determine the antecedent the pronoun refers to. In this German sentence, the antecedent *Europäische Zentralbank* dictates the feminine gender of the pronoun *sie*.

EN However, the European Central Bank (ECB) took an interest in it. *It* describes bitcoin as "the most successful virtual currency".

DE Dennoch hat die Europäische Zentralbank (EZB) Interesse hierfür gezeigt. *Sie* beschreibt Bitcoin als "die virtuelle Währung mit dem grössten Erfolg".

Figure 1: Example sentence illustrating how the translation of pronouns is ambiguous on a sentence level. Pronouns of interest are in italics, and the antecedents they refer to are underlined. Taken from WMT `newstest2013`.

It is unfortunate, then, that current NMT systems generally operate on the sentence level (Vaswani et al., 2017; Gehring et al., 2017; Hieber et al., 2017). Documents are translated sentence-by-sentence for practical reasons, such as line-based processing in a pipeline and reduced computational complexity. Furthermore, improvements of larger-context models over baselines in terms of document-level metrics such as BLEU or RIBES have been moderate, so that their computational overhead does not seem justified, and so that it is hard to develop more effective context-aware architectures and empirically validate them.

* Work performed prior to joining Amazon.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 61–72
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6307

To address this issue, we present an alternative way of evaluating larger-context models on a test set that allows to specifically measure a model's capability to correctly translate pronouns. The test suite consists of pairs of source and target sentences, in combination with contrastive translation variants (for evaluation by model scoring) and additional linguistic and contextual information (for further analysis). The resource is freely available.[1] Additionally, we evaluate several context-aware models that have recently been proposed in the literature on this test set, and extend existing models with parameter tying.

The main contributions of our paper are:

- We present a large-scale test set to evaluate the accuracy with which NMT models translate the English pronoun *it* to its German counterparts *es*, *sie* and *er*.
- We evaluate several context-aware systems and show how targeted, contrastive evaluation is an effective tool to measure improvement in pronoun translation.
- We empirically demonstrate the effectiveness of parameter tying in multi-encoder context-aware models.

Section 2 explains how our paper relates to existing work on context-aware models and the evaluation of pronoun translation. Section 3 describes our test suite. The context-aware models we use in our experiments are detailed in Section 4. We discuss our experiments in Section 5 and the results in Section 6.

2 Related Work

Two lines of work are related to our paper: research on context-aware translation (described in Section 2.1) and research on focused evaluation of pronoun translation (described in Section 2.2).

2.1 Context-Aware NMT Models

If the translation of a pronoun requires context beyond the current sentence (see the example in Figure 1), a natural extension of sentence-level NMT models is to condition the model prediction on this necessary context. In the following, we describe a number of existing approaches to making models "aware" of additional context.

The simplest possible extension is to translate units larger than sentences. Tiedemann and Scherrer (2017) concatenate each sentence with the sentence that precedes it, for the source side of the corpus or both sides. All of their models are standard sequence-to-sequence models built with recurrent neural networks (RNNs), since the method does not require any architectural change. Agrawal et al. (2018) use the same concatenation technique with a Transformer architecture (Vaswani et al., 2017), and experiment with wider context.

A number of works do propose changes to the NMT architecture. A common technique is to extend a standard encoder-decoder model by additional encoders for the context sentence(s), with a modified attention mechanism (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018). One aspect that differs between these works is the architecture of the encoder and attention. While Jean et al. (2017); Bawden et al. (2018) extend an RNN encoder-decoder with a second encoder that the decoder attends to, Voita et al. (2018) extend the Transformer architecture with an encoder that is attended to by the main encoder. Voita et al. (2018) also introduce parameter sharing between the main encoder and the context encoder, but do not empirically demonstrate its importance.

While the number of encoded sentences in the previous work is fixed, Wang et al. (2017); Maruf and Haffari (2018) explore the integration of variable-size context through a hierarchical architecture, where a first-level RNN reads in words to produce sentence vectors, which are then fed into a second-level RNN to produce a document summary.

Apart from differences in the architectures, related work varies in whether it considers source context, target context, or both (see Table 1 for an overview of language arcs and context types). Some work considers only source context, but for pronoun translation, target-side context is intuitively important for disambiguation, especially if the antecedent itself is ambiguous. In our evaluation, we therefore emphasize models that take into account both source and target context.

Our experiments are based on models from Bawden et al. (2018), who have released their source code.[2] We extend their models with parameter sharing, which was shown to be beneficial

[1] https://github.com/ZurichNLP/ContraPro

[2] https://github.com/rbawden/nematus

	Languages		Context types			
	source	target	source	target	preceding	following
Tiedemann and Scherrer (2017)	DE	EN	x	x	x	
Jean et al. (2017)	EN	FR/DE	x		x	
Wang et al. (2017)	ZH	EN	x		x	
Voita et al. (2018)	EN	RU	x		x	x
Bawden et al. (2018)	EN	FR	x	x	x	
Maruf and Haffari (2018)	FR/DE/ET	EN	x	x	x	
Agrawal et al. (2018)	EN	IT	x	x	x	x

Table 1: Overview of context-aware translation models in related work.

by Voita et al. (2018). Additionally, we consider a concatenative baseline, similar to Tiedemann and Scherrer (2017), and Transformer-based models (Voita et al., 2018).

2.2 Evaluation of Pronoun Translation

Pronouns can serve a variety of functions with complex cross-lingual variation (Guillou, 2016), and hand-picked, manually annotated test suites have been presented for the evaluation of pronoun translation (Guillou and Hardmeier, 2016; Isabelle et al., 2017; Bawden et al., 2018). While suitable for analysis, the small size of the test suites makes it hard to make statistically confident comparisons between systems, and the hand-picked nature of the test suites introduces biases.[3] To overcome these problems, we opted for a fully automatic approach to constructing a large-scale test suite.

Conceptually, our test set is most similar to the "cross-lingual pronoun prediction" task held at DiscoMT and WMT in recent years (Hardmeier et al., 2015; Guillou et al., 2016; Loáiciga et al., 2017): participants are asked to fill a gap in a target sentence, where gaps correspond to pronouns.

The first edition of the task focused on English→French, and it was found that local context (such as the verb group) was a strong signal for pronoun prediction. Hence, future editions only provided target-side lemmas instead of fully inflected forms, which makes the task less suitable to evaluate end-to-end neural machine translation systems, although such systems have been trained on the task (Jean et al., 2017).

Loáiciga et al. (2017) do not report on the proportion of intra-sentential and inter-sentential anaphora in their test set, but the two top-performing systems only made use of intra-sentential information. Our test suite focuses on allowing the comparison of end-to-end context-aware NMT systems, and we thus extract a large number of *inter-sentential anaphora*, with meta-data allowing for a focus on inter-sentential anaphora with a long distance between the pronoun and its antecedent. Our focus on evaluating end-to-end NMT systems also relieves us from having to provide annotated training sets, and reduces pressure to achieve balance and full coverage of phenomena.[4]

An alternative approach to automatically evaluate pronoun translation are reference-based methods that produce a score based on word alignment between source, translation output, and reference translation, and identification of pronouns in them, such as AutoPRF (Hardmeier and Federico, 2010) and APT (Miculicich Werlen and Popescu-Belis, 2017). Guillou and Hardmeier (2018) perform a human meta-evaluation and show substantial disagreement between reference-based metrics and human judges, especially because there often exist valid alternative translations that use different pronouns than the reference. Our test set, and our protocol of generating contrastive examples, is focused on selected pronouns to minimize the risk of producing contrastive examples that are actually valid translations.

3 Test set with contrastive examples

Contrastive evaluation requires a large set of suitable examples that involve the translation of pronouns. As additional goals, our test set is designed

[3] For example, all pronoun examples in the test suite by Bawden et al. (2018) require the previous target sentence for disambiguation, and thus do not reward models that condition on more than one sentence of context.

[4] For example, we do not consider cases where English *it* is translated into something other than a personal pronoun. While this would be a severe blind spot in a training set for pronoun prediction, the focused nature of our test suite does not impair the performance of end-to-end NMT systems on other phenomena.

Alignment	Frequency	Probability
it→es	255764	0.334
it→sie	64446	0.084
it→er	44543	0.058
it→ist	42614	0.055
it→Sie	26054	0.034
it→,	21037	0.027
it→das	17992	0.023
it→dies	11943	0.015
it→wird	11886	0.015
it→man	10539	0.013
it→ihn	7744	0.010

Table 2: Frequency and probability of alignments of *it* in the training data of our systems (all data from the WMT 2017 news translation task). Alignments are produced by a fast_align model.

to 1) focus on *hard* cases, so that it can be used as a benchmark to track progress in context-aware translation and 2) allow for fine-grained analysis.

Section 3.1 describes how we extract our data set. Section 3.2 explains how, given a set of contrastive examples, contrastive evaluation works.

3.1 Automatic extraction of contrastive examples from corpora

We automatically create a test set from the OpenSubtitles corpus (Lison and Tiedemann, 2016).[5] The goal is to provide a large number of difficult test cases where an English pronoun has to be translated to a German pronoun.

The most challenging cases are translating *it* to either *er, sie* or *es*, depending on the grammatical gender of the antecedent.[6] Not only is the translation of *it* ambiguous, there is also class imbalance in the training data (see Table 2). There is roughly a 30% probability that *it* is aligned to *es*,[7] which makes it difficult to learn to translate *er* and *sie*. We use parsing and automatic co-reference resolution to find translation pairs that satisfy our constraints.

To provide a basis for filtering with constraints, we tokenize the whole data set with the Moses tokenizer, generate symmetric word alignments with fast_align (Dyer et al., 2013), parse the English text with CoreNLP (Manning et al., 2014), parse the German text with ParZu (Sennrich et al., 2013) and perform coreference resolution on both sides. The coreference chains are obtained with the neural model of CoreNLP for English, and with CorZu for German (Tuggener, 2016), respectively.

Then we opt for high-precision, aggressive filtering, according to the following protocol: for each pair of sentences (e, f) in English and German, extract iff

- e contains the English pronoun *it*, and f contains a German pronoun that is third person singular (*er, sie* or *es*), as indicated by their part-of-speech tags;
- those pronouns are aligned to each other;
- both pronouns are in a coreference chain;
- their nominal antecedents in the coreference chain are aligned on word level.

This removes most candidate pairs, but is necessary to overcome the noise introduced by our preprocessing pipeline, most notably coreference resolution. From the filtered set, we create a balanced test set by randomly sampling 4000 instances of each of the three translations of *it* under consideration (*er*, *sie*, *es*). We do not balance antecedent distance. See Table 4 for the distribution of pronoun pairs and antecedent distance in the test set.

For each sentence pair in the resulting test set, we introduce *contrastive translations*. A contrastive translation is a translation variant where the correct pronoun is swapped with an incorrect one. For an example, see Table 3, where the pronoun *it* in the original translation corresponds to *sie* because the antecedent *bat* is a feminine noun in German (*Fledermaus*). We produce wrong translations by replacing *sie* with one of the other pronouns (*er*, *es*).

Note that, by themselves, these contrastive translations are grammatically correct if the antecedent is outside the current sentence. The test set also contains pronouns with an antecedent in the same sentence (antecedent distance 0). Those examples do not require any additional context

[5] http://opus.nlpl.eu/OpenSubtitles2016.php

[6] The pronouns *he* and *she* usually refer to a person in English, and since persons do not change gender in the translation, we assume that learning the correspondences *he* → *er* and *she* → *sie* does not present a challenge for a model. Cases where *he* or *she* refer to a noun that is not a person are possible, but extremely rare.

[7] Note that these statistics include non-referential uses of *it*, that we exclude from our testset.

source:	*It could get tangled in your hair.*
reference:	***Sie** könnte sich in deinem Haar verfangen.*
contrastive:	***Er** könnte sich in deinem Haar verfangen.*
contrastive:	***Es** könnte sich in deinem Haar verfangen.*
antecedent en:	a bat
antecedent de:	eine Fledermaus (f.)
antecedent distance :	1

Table 3: Example sentence pair with contrastive translations. An antecedent distance of 1 means that the antecedent is in the immediately preceding sentence.

for disambiguation and we therefore expect the sentence-level baseline to perform well on them.

We take extra care to ensure that the resulting contrastive translations are grammatically correct, because ungrammatical sentences are easily dismissed by an NMT system. For instance, if there are any possessive pronouns (such as *seine*) in the sentence, we also change their gender to match the personal pronoun replacement.

The German coreference resolution system does not resolve *es* because most instances of *es* in German are either non-referential forms, or they refer to a clause instead of a nominal antecedent. We limit the test set to nominal antecedents, as these are the only ambiguous cases with respect to translation. For this reason, we have to rely entirely on the English coreference links for the extraction of sentence pairs with *it→es*, as opposed to pairs with *it→er* and *it→sie* where we have coreference chains in both languages.[8]

Our extraction process respects document boundaries, to ensure we always search for the right context. We extract additional information from the annotated documents, such as the distance (in sentences) between pronouns and their antecedents, the document of origin, lemma, morphology and dependency information if available.

3.2 Evaluation by scoring

Contrastive evaluation is different from conventional evaluation of machine translation in that it does not require any translation. Rather than testing a model's ability to translate, it is a method to test a model's ability to *discriminate* between given good and bad translations.

distance	*it→es*	*it→er*	*it→sie*	total
0	872	736	792	2400
1	1892	2577	2606	7075
2	631	459	420	1510
3	274	167	132	573
>3	331	61	50	442
total	4000	4000	4000	12000

Table 4: Test set frequencies of pronoun pairs and antecedent distance (measured in sentences).

We exploit the fact that NMT systems are in fact language models of the target language, conditioned on source text. Like language models, NMT systems can be used to compute a model score (the negative log probability) for an existing translation. Contrastive evaluation, then, means to compare the model score of two pairs of inputs: $(actual\ source,\ reference\ translation)$ and $(actual\ source,\ contrastive\ translation)$. If the model score of the actual reference translation is higher, we assume that this model can detect wrong pronoun translations.

However, this does *not* mean that systems actually produce the reference translation when given the source sentence for translation. An entirely different target sequence might rank higher in the system's beam during decoding. The only conclusion permitted by contrastive evaluation is whether or not the reference translation is more probable than a contrastive variant.

If the model score of the reference is indeed higher, we refer to this outcome as a "correct decision" by the model. The model's decision is only correct if the reference translation has a higher score than any contrastive translation. In our evaluation, we aggregate model decisions on

[8] There are some cases where the antecedent is listed as *it* in the test set. This is our fallback behaviour if the coreference chain does not contain any noun. In that case, we do not know the true antecedent.

the whole test set and report the overall percentage of correct decisions as accuracy.

During scoring, the model is provided with reference translations as target context, while during translation, the model needs to predict the full sequence. It is an open question to what extent performance deteriorates when context is itself predicted, and thus noisy. We highlight that the same problem arises for sentence-level NMT, and has been addressed with alternative training strategies (Ranzato et al., 2015).

4 Context-Aware NMT Models

This section describes several context-aware NMT models that we use in our experiments. They fall into two major categories: models based on RNNs and models based on the Transformer architecture (Vaswani et al., 2017). We experiment with additional context on the source side and target side.

4.1 Recurrent Models

We consider the following recurrent baselines:

baseline Our baseline model is a standard bidirectional RNN model with attention, trained with Nematus. It operates on the sentence level and does not see any additional context. The input and output embeddings of the decoder are tied, encoder embeddings are not.

concat22 We concatenate each sentence with one preceding sentence, for both the source and target side of the corpus. Then we train on this new data set without any changes to the model architecture. This very simple method is inspired by Tiedemann and Scherrer (2017).

The following models are taken, or slightly adapted, from Bawden et al. (2018). For this reason, we give only a very short description of them here and the reader is referred to their work for details.

s-hier A multi-encoder architecture with hierarchical attention. This model has access to one additional context: the previous source sentence. It is read by a separate encoder, and attended to by an additional attention network. The output of the resulting two attention vectors is combined with yet another attention network.

s-t-hier Identical to *s-hier*, except that it considers two additional contexts: the previous source sentence and previous target sentence. Both are read by separate encoders, and sequences from all encoders are combined with hierarchical attention.

s-hier-to-2 The model has an additional encoder for source context, whereas the target side of the corpus is concatenated, in the same way as for *concat22*. This model achieved the best results in Bawden et al. (2018).

For each variant, we also introduce and test weight tying: we share the parameters of embedding matrices between encoders that read the same kind of text (source or target side).

4.2 Transformer Models

All remaining models are based on the Transformer architecture (Vaswani et al., 2017). A Transformer avoids recurrence completely: it follows an encoder-decoder architecture using stacked self-attention and fully connected layers for both the encoder and decoder.

baseline A standard context-agnostic Transformer. All model parameters are identical to a *Transformer-base* in Vaswani et al. (2017).

concat22 A simple concatentation model where only the training data is modified, in the same way as for the recurrent *concat22* model.

concat21 Trained on data where the preceding sentence is concatenated to the current one only on the source side. This model is also taken from Tiedemann and Scherrer (2017).

Voita et al. (2018) A more sophisticated context-aware Transformer that uses source context only. It has a separate encoder for source context, but all layers except the last one are shared between encoders. A source and context sentence are first encoded independently, and then a single attention layer and a gating function are used to produce a context-aware representation of the source sentence. Such restricted interaction with context is shown to be beneficial for analysis of contextual phenomena captured by the model. For details the reader is referred to their work.

5 Experiments

We train all models on the data from the WMT 2017 English→German news translation shared task ($\sim$ 5.8 million sentence pairs). These corpora do not have document boundaries, therefore a small fraction of sentences will be paired with wrong context, but we expect the model to be robust against occasional random context (see also Voita et al. 2018). Experimental setups for the RNN and Transformer models are different, and we describe them separately.

All RNN-based models are trained with Nematus (Sennrich et al., 2017). We learn a joint BPE model with 89.5k merge operations (Sennrich et al., 2016). We train shallow models with an embedding size of 512, a hidden layer size of 1024 and layer normalization. Models are trained with Adam (Kingma and Ba, 2015), with an initial learning rate of 0.0001. We apply early stopping based on validation perplexity. The batch size for training is 80, and the maximum length of training sequences is 100 (if input sentences are concatenated) or 50 (if input lines are single sentences).

For our Transformer-based experiments, we use a custom implementation and follow the hyperparameters from Vaswani et al. (2017); Voita et al. (2018). Systems are trained on lowercased text that was encoded using BPE (32k merge operations). Models consist of 6 encoder and decoder layers with 8 attention heads. The hidden state size is 512, the size of feedforward layers is 2048.

Model performance is evaluated in terms of BLEU, on `newstest2017`, `newstest2018` and all sentence pairs from our pronoun test set. We compute scores with SacreBLEU (Post, 2018).[9] Evaluation with BLEU is done mainly to control for overall translation quality.

To evaluate pronoun translation, we perform contrastive evaluation and report the accuracy of models on our contrastive test set.

6 Evaluation

The BLEU scores in Table 5 show a moderate improvement for most context-aware systems. This suggests that the architectural changes for the context-aware models do not degrade overall translation quality. The contrastive evaluation on our test set on the other hand shows a clear increase in the accuracy of pronoun translation: The best model *s-hier-to-2.tied* achieves a total of +16 percentage points accuracy on the test set over the baseline, see Table 6.

Table 7 shows that context-aware models perform better than the baseline when the antecedent is outside the current sentence. In our experiments, all context-aware models consider one preceding sentence as context. The evaluation according to the distance of the antecedent in Table 8 confirms that the subset of sentences with antecedent distance 1 benefits most from the tested context-aware models (up to +20 percentage points accuracy). However, we note two surprising patterns:

- For inter-sentential anaphora, the performance of all systems, including the baseline, improves with increasing antecedent distance.
- Context-aware systems that consider one preceding sentence also improve on intra-sentential anaphora, and on pronouns whose antecedent is outside the context window.

The first observation can be explained by the distribution of German pronouns in the test set. The further away the antecedent, the higher the percentage of *it→es* cases, which are the majority class, and thus the class that will be predicted most often if evidence for other classes is lacking. We speculate that this is due to our more permissive extraction heuristics for *it→es*.

We attribute the second observation to the existence of coreference chains where the preceding sentence contains a pronoun that refers to the same nominal antecedent as the pronoun in the current sentence. Consider the example in Table 9: The nominal antecedent of *it* in the current sentence is *door*, *Tür* in German with feminine gender. The nominal antecedent occurs two sentences before the current sentence, but the German sentence in between contains the pronoun *sie*, which is a useful signal for the context-aware models, even though they cannot know the nominal antecedent.

Note that only models aware of target-side context can benefit from such circumstances: The *s-hier* models as well as the Transformer model by (Voita et al., 2018) only see source side context, which results in lower accuracy if the distance to the antecedent is >1, see Table 8.

While such coreference chains complicate the interpretation of the results, we note that improvements on inter-sentential anaphora with antecedent distance > 1 are relatively small (compared to distance 1), and that performance is still relatively poor (especially for the minority classes *er* and *sie*). We encourage evaluation of wider-context models on this subset, which is still large thanks to the size of the full test set.

Regarding the comparison of different context-aware architectures, our results demonstrate the

[9] Our (cased) SacreBLEU signature is `BLEU+c.mixed+l.en-de+#.1+s.exp+t.wmt{17,18}+tok.13a+v.1.2.10`.

	newstest2017		newstest2018		pronoun set	
	cased	uncased	cased	uncased	cased	uncased
baseline	23.0	23.7	33.7	34.2	19.4	19.9
concat22	23.8	24.4	**34.5**	35.0	**20.2**	20.8
independent encoders						
s-hier	23.5	24.0	33.5	34.0	18.9	19.5
s-hier-to-2	23.8	24.3	34.2	34.8	19.2	19.7
s-t-hier	23.1	23.6	33.1	33.6	19.3	20.0
with weight tying						
s-hier.tied	23.6	24.1	33.7	34.2	19.7	20.3
s-hier-to-2.tied	**24.2**	24.8	34.1	34.7	20.1	20.7
s-t-hier.tied	23.5	24.0	33.9	34.5	19.4	20.0
Transformer-based models						
baseline	-	24.6	-	35.4	-	21.1
concat21	-	24.8	-	35.3	-	**21.8**
concat22	-	24.4	-	36.0	-	21.3
(Voita et al., 2018)	-	**25.3**	-	**36.5**	-	21.7

Table 5: English→German BLEU scores on newstest2017, newstest2018 and all sentence pairs from our pronoun test set. Case-sensitive and case-insensitive (uncased) scores are reported. Higher is better, and the best scores are marked in bold.

		reference pronoun		
	total	*es*	*er*	*sie*
baseline	0.44	0.85	0.17	0.31
concat22	0.53	0.84	0.32	0.42
independent encoders				
s-hier	0.43	0.80	0.20	0.29
s-hier-to-2	0.55	0.84	0.41	0.40
s-t-hier	0.52	0.88	0.32	0.36
with weight tying				
s-hier.tied	0.47	0.85	0.30	0.26
s-hier-to-2.tied	**0.60**	0.87	**0.45**	**0.48**
s-t-hier.tied	0.56	0.86	0.39	0.42
Transformer-based models				
baseline	0.47	0.81	0.22	0.38
concat21	0.48	0.88	0.26	0.31
concat22	0.49	**0.91**	0.20	0.36
(Voita et al., 2018)	0.49	0.84	0.23	0.39

Table 6: Accuracy on contrastive test set (N=4000 per pronoun) with regard to reference pronoun.

	antecedent location	
	intrasegmental	external
baseline	0.57	0.41
concat22	0.58	0.51
independent encoders		
s-hier	0.58	0.39
s-hier-to-2	0.63	0.53
s-t-hier	0.52	0.52
with weight tying		
s-hier.tied	0.56	0.45
s-hier-to-2.tied	0.65	**0.58**
s-t-hier.tied	0.57	0.55
Transformer-based models		
baseline	0.70	0.41
concat21	0.67	0.44
concat22	0.56	0.47
(Voita et al., 2018)	**0.75**	0.43

Table 7: Accuracy on contrastive test set with regard to antecedent location (within segment vs. outside segment).

	antecedent distance				
	0	1	2	3	>3
baseline	0.57	0.38	0.47	0.52	0.67
concat22	0.58	0.50	0.51	0.51	0.69
independent encoders					
s-hier	0.58	0.36	0.42	0.46	0.61
s-hier-to-2	0.63	0.51	0.54	0.60	0.70
s-t-hier	0.52	0.49	**0.57**	**0.61**	0.71
with weight tying					
s-hier.tied	0.56	0.43	0.46	0.49	0.67
s-hier-to-2.tied	0.65	**0.58**	0.55	0.55	**0.75**
s-t-hier.tied	0.57	0.54	0.56	0.59	0.72
Transformer-based models					
baseline	0.70	0.38	0.45	0.49	0.65
concat21	0.67	0.42	0.45	0.47	0.66
concat22	0.56	0.44	0.53	0.54	0.74
(Voita et al., 2018)	**0.75**	0.39	0.48	0.54	0.66

Table 8: Accuracy on contrastive test set with regard to antecedent distance of antecedent (in sentences).

source sentence with antecedent	*What's with the door?*
target sentence with antecedent	*Was ist mit der Tür?*
source context	***It** won't open.*
reference context	***Sie** geht nicht auf.*
source sentence	*- Is **it** locked?*
reference sentence	*- Ist **sie** abgeschlossen?*
contrastive 1	*- Ist **er** abgeschlossen?*
contrastive 2	*- Ist **es** abgeschlossen?*

Table 9: Example where 1) antecedent distance is >1 and 2) the context given contains another pronoun as an additional hint.

effectiveness of parameter sharing between the main encoder (or decoder) and the contextual encoder. We observe an improvement of 5 percentage points from *s-hier-to-2* to *s-hier-to-2.tied*, and 4 percentage points from *s-t-hier* to *s-t-hier.tied*. Context encoders introduce a large number of extra parameters, while inter-sentential context is only relevant for a relatively small number of predictions. We hypothesize that the training signal is thus too weak to train a strong contextual encoder in an end-to-end fashion without parameter sharing. Our results also confirm the finding by Bawden et al. (2018) that multi-encoder architectures, specifically *s-hier-to-2(.tied)*, can outperform a simple concatenation system in the translation of coreferential pronouns.

The Transformer-based models perform strongest on pronouns with intra-segmental antecedent, outperforming the recurrent baseline by 9–18 percentage points. This is likely an effect of increased model depth and the self-attentional architecture in this set of experiments. The model by (Voita et al., 2018) only uses source context, and outperforms the most comparable RNN system, *s-hier.tied*. However, the Transformer-based *concat22* slightly underperforms the RNN-based *concat22*, and we consider it future research how to better exploit target context with Transformer-based models.

7 Conclusions

We present a large-scale test suite to specifically test the capacity of NMT models to translate pronouns correctly. The test set contains 12,000 difficult cases of pronoun translations from English *it* to its German counterparts *er, sie* and *es*, extracted automatically from OpenSubtitles (Lison and Tiedemann, 2016).

We evaluate recently proposed context-aware models on our test set. Even though the increase in BLEU score is moderate for all context-aware models, the improvement in the translation of pronouns is considerable: The best model (*s-hier-to-2.tied*) achieves a +16 percentage points gain in accuracy over the baseline.

Our experiments confirm the importance of careful architecture design, with multi-encoder architectures outperforming a model that simply concatenates context sentences. We also demonstrate the effectiveness of parameter sharing between encoders of a context-aware model.

We hope the test set will prove useful for empirically validating novel architectures for context-aware NMT. So far, we have only evaluated models that consider one sentence of context, but the nominal antecedent is more distant for a sizable proportion of the test set, and the evaluation of variable-size context models (Wang et al., 2017; Maruf and Haffari, 2018) is interesting future work.

Acknowledgements

We are grateful to the Swiss National Science Foundation (SNF) for supporting the project CoNTra (grant number 105212_169888).

References

Ruchit Agrawal, Turchi Marco, and Negri Matteo. 2018. Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides.

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. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *NAACL 2018*, New Orleans, USA.

Chris Dyer, 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*, pages 644–648. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2018. Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 wmt shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused mt and cross-lingual pronoun prediction: Findings of the 2015 discomt shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? *ArXiv e-prints*.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 54–57, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, San Diego, USA. Ithaca, NY.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 discomt shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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, ACL*, pages 1275–1284, Melbourne, Australia.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT)*, EPFL-CONF-229974. Association for Computational Linguistics (ACL).

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, page 5998–6008. Curran Associates, Inc.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *ACL 2018*, Melbourne, Australia.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.