Analysing concatenation approaches to document-level NMT in two different domains

Yves Scherrer[1] **Jörg Tiedemann**[1] **Sharid Loáiciga**[2]
[1]Department of Digital Humanities, University of Helsinki
[2]CLASP, Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg
yves.scherrer@helsinki.fi jorg.tiedemann@helsinki.fi
sharid.loaiciga@gu.se

Abstract

In this paper, we investigate how different aspects of discourse context affect the performance of recent neural MT systems. We describe two popular datasets covering news and movie subtitles and we provide a thorough analysis of the distribution of various document-level features in their domains. Furthermore, we train a set of context-aware MT models on both datasets and propose a comparative evaluation scheme that contrasts coherent context with artificially scrambled documents and absent context, arguing that the impact of discourse-aware MT models will become visible in this way. Our results show that the models are indeed affected by the manipulation of the test data, providing a different view on document-level translation quality than absolute sentence-level scores.

1 Introduction

Shortly after the change of paradigm in Machine Translation (MT) from statistical to neural architectures, the interest in discourse phenomena flourished again. This is not by chance, as neural models can embed larger text spans into contextual representations and can be set up to learn relevant features from the raw data to produce better translations.

It is still unclear though how the impact of discourse on MT quality should be evaluated and analyzed. On one side, it is difficult to pinpoint particular contextual features that neural MT (NMT) models are picking up. On the other, it is difficult to judge good translations purely in terms of discourse features. In this paper, we investigate the discourse-related biases in data. Our contributions are twofold:

- we provide a thorough analysis of two popular machine translation datasets in terms of document-level features,
- we train different context-aware MT models (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Maruf et al., 2019; Junczys-Dowmunt, 2019) on the two datasets and evaluate them using a comparative setup with artificially scrambled data.

As discourse properties of the data, we consider pronouns and coreference chains, connectives, and negation. For the evaluation of translation quality and the influence of document-level context, we contrast context-aware models at test time with (1) clean coherent text, (2) incoherent input and (3) zero-context input.[1] For the second type, we scramble sentences and insert document boundaries at arbitrary positions in the test data. For the third approach, we add document boundaries after each test instance. This setup provides a cheap way of testing the influence of contextual information on translation performance that can be measured in common ways, for example, facilitating automatic evaluation metrics such as BLEU or METEOR.

2 Related work

2.1 Discourse

Research about discourse and MT has shifted from explicitly enhancing systems with discourse knowledge to evaluating how much the systems have learned specific discourse features through different resources, test suites being a popular one (cf. Sim Smith, 2017; Popescu-Belis, 2019). Throughout, however, particular discourse phenomena are consistently targeted, as they are indeed indicators of globally good, cohesive and coherent texts. Pronouns (Hardmeier and Federico, 2010; Guillou, 2012; Hardmeier et al., 2013;

[1] *Context* here refers to text outside of the sentence to be translated.

Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 51–61
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics

Guillou and Hardmeier, 2016; Müller et al., 2018; Guillou et al., 2018) have been largely at the center of attention, and more recently the translation of pronouns in the context of their coreferential chains has been looked at (Lapshinova-Koltunski and Hardmeier, 2017; Voita et al., 2018; Lapshinova-Koltunski et al., 2019). Other devices studied are verbal tenses (Gong et al., 2012; Loáiciga et al., 2014; Ramm and Fraser, 2016) and connectives (Meyer et al., 2012; Meyer and Popescu-Belis, 2012), although not using neural models. Motivated by approximating the ability of systems to grasp more abstract properties related to coherence, ambiguous words have also been targeted (Rios Gonzales et al., 2017; Bawden et al., 2018; Rios et al., 2018), as well as ellipsis (Voita et al., 2019). Last, negation (Fancellu and Webber, 2015) is a rather understudied phenomenon, but like pronouns and their antecedents, the scope of the negation can be in a different sentence.

In this paper we investigate these features in the training data and assess translation using standard automatic metrics and a data scrambling strategy.

2.2 Context-aware NMT

Tiedemann and Scherrer (2017) present a simple approach to context-aware NMT: instead of training the model on pairs of single source and target sentences, they add sentences from the left context to the sentence to be translated, either only on the source side or both on source and target sides. These models are evaluated on a German–English corpus extracted from OpenSubtitles, and the best results are obtained with two source sentences and one target sentence. Agrawal et al. (2018) extend these experiments by considering additional contexts. They evaluate their work on the IWSLT 2017 dataset for English–Italian, which consists of transcripts of TED talks.

In 2019, the WMT conference featured for the first time a document-level translation task for English–German (Barrault et al., 2019). One of the best-performing systems (Junczys-Dowmunt, 2019) is based on a similar idea: all sentences of a document are concatenated and translated as a whole. Documents whose length exceeds the maximum sequence length defined by the model are simply split.

The approaches outlined above, which we refer to as "concatenation models", do not require any change to the NMT model architecture. Other recent work explores the feasability of extending NMT models to make them context-aware. A common approach is to use additional encoders for the context sentence(s) with a modified attention mechanism (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018). Another technique (Miculicich et al., 2018; Maruf et al., 2019) explores the integration of context through a hierarchical architecture which models the contextual information in a structured manner using word-level and sentence-level abstractions.

The different models have been evaluated on different language pairs and different datasets. In this paper, we focus on a single language pair, English–German (in both directions), and on two textual domains: news translation and movie subtitles translation. For the news translation task (denoted as *WMT*) we rely on the established setup of WMT 2019[2] with the Newstest2018 data as our dedicated test set. For the movie subtitles (referred to as *OST*), we use data from the OpenSubtitles corpus released on OPUS[3] with our own split into training, development and test data. More details about the data and our setup will be given in the following section.

3 Two datasets for English–German document-level translation

Different text genres and types exhibit different types of discourse-level properties. The choice of training corpus therefore determines what features a NMT model can potentially learn, and the choice of test corpus determines which features can be reliably evaluated. Our experiments are based on two datasets that cover the same language pair, but very different textual characteristics.

The **OST** dataset is built from the English–German part of the publicly available OpenSubtitles2016 corpus (Lison and Tiedemann, 2016). Of the 16,910 movies and TV series in the collection, 16,510 are used for training, and 4 each are held out for development and testing purposes. Each movie is considered a single document. It corresponds to the dataset used in Tiedemann and Scherrer (2017). General properties of this dataset can be found in Table 1.

The **WMT** dataset comprises the subset of corpora allowed at the WMT 2019 news translation

[2] See http://www.statmt.org/wmt19/translation-task.html.

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

Corpus	Documents	Sentences	Sents/Doc	Tokens DE	Tokens EN	Tokens/Sent
OST Train	16,510	13,544k	820	104,447k	111,729k	8.0
OST Valid	4	5k	1249	41k	43k	8.4
OST Test	4	5k	1249	38k	47k	8.4
WMT Train	583,358	12,690k	22	259,384k	276,401k	21.1
WMT Valid	236	5k	22	106k	111k	21.1
WMT Test	122	3k	25	64k	68k	21.9

Table 1: General characteristics of the two datasets. Tokens/Sent values are averaged over the DE and EN tokens.

task which contains document boundaries. The training set includes parallel data from the Europarl v9, NewsCommentary v14, and Rapid2019 collections. We select the Newstest2015 and Newstest2016 corpora as our validation set and the Newstest2018 corpus as our test set. General properties of this dataset can be found in Table 1.

Table 1 shows that the two datasets are comparable in terms of sentence numbers.[4] However, the documents in OST are up to 50 times larger than those in WMT (cf. column *Sents/Doc*). On the other hand, WMT sentences are more than twice as long than OST sentences (cf. column *Tokens/Sent*), which is in line with our expectations.

A third dataset based on transcripts of TED talks (Cettolo et al., 2012), has also been used for document-level translation (Agrawal et al., 2018). We do not consider this dataset for training due to its smaller size, but use the PROTEST test suite, which is based on this corpus, for evaluation (Guillou and Hardmeier, 2016; Guillou et al., 2018).

3.1 Discourse-level properties

In recent literature, various linguistic features have been identified to contribute to document-level coherence and cohesion. In this section, we assess the two datasets in order to estimate their suitability and difficulty for document-level translation. We investigate the following phenomena:

Pronouns: We first extract a list of pronouns per language by tagging the training corpora with SpaCy[5], extracting the tokens labeled as PRON and manually cleaning the resulting list (cf. Table 7). Then, the frequency of pronouns is computed independently for English and German.

The results in Table 2 show that about every 10th word of the OST corpus is a pronoun, whereas pronouns are three to four times rarer in the WMT corpus.[6] This divergence is to be expected, as OST consists mainly of dialogues.

Not all pronouns are intrinsically hard to translate. Therefore, we also examine how many **ambiguous pronouns** occur in the corpora. To this end, the English and German corpora are word-aligned using Eflomal (Östling and Tiedemann, 2016) and for each source pronoun (as defined in the list extracted previously), the target pronouns are retrieved. If this list contains at least two words totalling each at least 10% of occurrences, we consider the source pronoun as ambiguous (cf. Table 7). This feature is computed separately for both translation directions.

On average, about half of the pronoun occurrences are ambiguous, with most ambiguities concerning case (e.g. *me* translating both to accusative *mich* and dative *mir*). The English pronouns in the OST dataset deviate from this tendency, mainly because of the prevalence of *you*: this pronoun is ambiguous both in terms of number and politeness and can be translated as *du*, *ihr*, or *Sie* (see also Sennrich et al., 2016).

Connectives: As part of their *Accuracy of Connective Translation* metric, Hajlaoui and Popescu-Belis (2013) provide a list of eight ambiguous English connectives and their German translations. We count the number of sentence pairs that contain both an English connective and one of its German translations, regardless of its associated sense.

Ambiguous connectives show an inverse frequency distribution compared to pronouns: they are about ten times as frequent in WMT than in OST. This divergence can again be attributed to genre differences.

[4] By sentences, we mean the lines obtained by the sentence alignment process.

[5] spacy.io

[6] The numbers for German are higher because the pronoun list contains more relative and demonstrative pronouns than the English one, as a result of annotation differences in the SpaCy training corpora.

Corpus	Pronouns		Ambiguous pronouns		Ambiguous connectives	Negations		Negation discrep.	Coreference chains		Cross-sent. pron. coref.	
	DE	EN	DE	EN	DE–EN	DE	EN	DE–EN	DE	EN	DE	EN
OST Train	106.0	97.0	44.1	71.1	5.0	151.6	162.8	57.1	290.5	148.3	67.2	44.5
OST Valid	104.7	92.7	49.9	73.0	6.2	165.5	171.5	65.6	346.1	167.5	70.2	46.4
OST Test	101.1	99.3	53.0	69.7	5.8	148.9	191.9	75.0	292.5	178.8	66.8	46.9
WMT Train	36.1	20.0	20.1	13.5	60.2	176.1	176.2	19.6	670.3	495.3	91.9	80.6
WMT Valid	44.2	29.6	24.6	20.8	62.5	182.1	177.2	23.8	693.5	544.2	111.5	97.6
WMT Test	44.0	25.8	25.9	20.0	58.3	167.4	169.1	18.3	726.8	535.0	115.4	99.7
	per thousand tokens				per thousand lines							

Table 2: Discourse-level features in the OST and WMT datasets. Coreference values were computed on a subset of the training corpora.

Negations: We establish a list of sentential and nominal negation words for both languages (cf. Table 7) and count the number of sentences that contain at least one negation word. We also count **negation discrepancies**, i.e. aligned sentence pairs where a negation was identified in one language but not in the other.

While the overall frequencies of negations are similar in both corpora, there are significantly more discrepancies in the OST dataset. These can be ascribed to two factors: free translation (a negation can be paraphrased with expressions such as *fail to*, *doubt if*, etc.), and sentence alignment errors.

Coreference chains: We assume that a large amount of pronouns, connectives and negations do not require access to large contexts for their correct translation, either because they are unambiguous or because the current sentence is sufficient for their disambiguation. To corroborate this assumption, we annotate the English corpora with the Stanford CoreNLP coreference resolver (Manning et al., 2014; Clark and Manning, 2016) and the German corpora with the CorZu coreference resolver (Tuggener, 2016).[7]

We first report the numbers of coreference chains identified by the resolvers. These numbers are hard to compare across languages due to different performance levels of the two resolvers, and translationese factors such as explicitation. However, they confirm the intuition that news text contains more referring entities than movie dialogues.[8]

Second, we count **cross-sentential pronominal coreference chains**, i.e. chains that span at least two sentences, contain at least one third-person pronoun and at least two different mention strings. The results suggest that about every 10th line of the WMT dataset and about every 20th line of the OST dataset contains a pronoun that requires access to the context for its correct translation. Given the overall training data sizes, NMT models should thus be able to pick up this signal.

Overall, the examined discourse-level features show consistent patterns across the training, validation and test sets. This was not necessarily expected for the WMT corpus, whose training set stems from a wide variety of sources.[9]

Three other discourse-level features could have been analyzed as well: We did not include verbal tenses, as we do not expect them to be particularly problematic for the German–English language pair. Likewise, we did not include measures for lexical consistency (Carpuat and Simard, 2012), as this was already reported to be handled well in SMT. Finally, we did not include ellipsis (Voita et al., 2019) as we found it difficult to detect and not very relevant for German.

4 Context-aware MT models

In this paper, our main focus lies on concatenation models as one of the most straightforward and successful approaches to document-level NMT. We train various concatenation models on both datasets and for both translation directions in order to perform a systematic study on this setup.

[7] Due to slow performance, we could only analyze 13% of the English OST, 5% of the English WMT and 5% of the German WMT training sets. We nevertheless believe that the reported proportions are representative of the entire dataset.

[8] Note also that the WMT dataset may benefit from higher recall as the coreference resolution pipelines are typically trained on newswire data.

[9] For the MT training, we shuffle the datasets keeping documents and document boundaries intact.

Inspired by Agrawal et al. (2018), we name the configurations according to the following schema:

$$i\textbf{Prev + Curr + }j\textbf{Next} \rightarrow k\textbf{Prev + Curr}$$

where i denotes the number of previous sentences on the source side, j the number of following sentences on the source side, and k the number of previous sentences on the target side. In all models, only the current sentence is evaluated. The following configurations are tested:

- Curr → Curr (baseline)
- 1Prev + Curr → Curr
- 1Prev + Curr + 1Next → Curr
- 2Prev + Curr → Curr
- 1Prev + Curr → 1Prev + Curr
- 1Prev + Curr + 1Next → 1Prev + Curr

Several discourse-level properties, among which most prominently pronoun gender, also depend on the previously generated output in the target language. Therefore, we also include an oracle variant where we the reference translation of the previous sentence (instead of its source) is fed to the system:

- 1PrevTarget + Curr → Curr

Furthermore, we also train fixed window models as in Junczys-Dowmunt (2019):

- 100T → 100T: A model that sees chunks of at most 100 tokens (after subword encoding) on either source and target side.
- 250T → 250T: A model that sees chunks of at most 250 tokens (after subword encoding) on either source and target side.

Note that these chunks are not produced using a sliding window but rather break documents at arbitrary positions unless they are less than the maximum size in length. We adopt the same annotation scheme as proposed in the original approach, marking segment and document boundaries with special symbols for document-internal breaks and continuations. We never break sentences from the original alignment into pieces, which would negatively affect the model and complicate the alignment of training examples.

The chosen chunk lengths seem very small, especially when considering subword units. Table 3 lists some basic statistics that demonstrate the effect of the chunking approach. We can see that even 100-token windows create reasonably large units that combine context beyond sentence boundaries. For the WMT dataset with larger sentences, we observe an average of almost 3 joined segments per chunk. For the subtitle data, the situation is much more extreme: most segments are very short and a 100-token window corresponds to about 10 segments. Hence, this approach yields a substantial increase of contextual information compared to the baseline.

Window size	Chunks	Sents/chunk
OST training data:		
100 tokens	1 282 985	10.6
250 tokens	496 207	27.3
WMT training data:		
100 tokens	4 286 535	3.0
250 tokens	1 729 601	7.3

Table 3: Basic statistics of fixed-size windows data.

Junczys-Dowmunt (2019) suggested to use even larger chunks, but that did not seem to work well in our current settings. Already the second model with a maximum of 250 tokens did not converge to any reasonable result when trained from scratch. We tried to address this problem by initialising the larger model with a pre-trained 100-token model but this approach did not lead to satisfactory results either. Therefore, we exclude all models larger than 100 tokens from our discussions below.

All models are based on the standard Transformer architecture and were trained with MarianNMT (Junczys-Dowmunt et al., 2018). For the WMT EN→DE models, we added 10.3M lines of backtranslations. These backtranslations consisted of German news documents (News2018) translated to English with a sentence-level model; document boundaries were kept intact. We did not include backtranslations for the opposite translation direction to investigate their impact on discourse-level translation.

Our experiments with recently proposed hierarchical attention networks for document-level NMT, in particular Miculicich et al. (2018) and Maruf et al. (2019), either underperformed or could not cope with the data sizes and document lengths of our training sets. For comparison, we nevertheless report results of a selective attention (Maruf et al., 2019) model for the WMT

EN→DE task. This model has to be trained in a two-step procedure: (1) a standard sentence-level model is trained on all the training data and, (2) a document-level model is trained on top of the sentence-level model that adds the inter-sentential information from the surrounding context using the attentive connections of the extended network. We focused on source-side attention for the wider context and did not explore further setups due to computational costs and unsatisfactory baseline results. Otherwise, we use the standard settings recommended in the released software.

5 Evaluation

Each system is evaluated on the respective test set using the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) metrics. In particular, we evaluate each of them on three variants of the test set:

Consistent context: the context sentences of the test set are appended in their natural order, as they appear in the data.

Inconsistent context: the test set is shuffled such that the context sentences are random.

No context: each sentence of the test set is considered its own document, so no contextual information is made available.

This setup allows us to check whether observed improvements are due to the additional context or to other factors.[10] A good context-aware system should perform best with consistent context and worst with inconsistent context.

Note that the concatenation models need some special treatment at test time. The sliding window approaches need to be post-processed in order to remove non-relevant parts of the translation in all cases where we train models with extended target language content. For simplicity, we rely on the segment separation tokens that are produced in translation similar to the ones seen during training. We have found this approach to be very robust, in the sense that the models reliably learn to place them at appropriate positions.

For the non-sliding window approaches with fixed maximum size, sentence splitting is not as straightforward and requires some additional treatment. Segments are also separated by separation tokens but we realized that they do not necessarily match with the segment boundaries in the reference data even though the original paper suggests that this should be rather stable (Junczys-Dowmunt, 2019). This is especially fatal if the number of segments does not match. Therefore, we apply standard sentence alignment based on length-correlation and lexical matches using hunalign (Varga et al., 2005) to link the system output to the reference translations. The reported results from the fixed-size models are based on this approach.

5.1 Generic translation metrics

We report BLEU and METEOR scores for all our experiments in Tables 4 and 5. The results and significance tests were computed using *MultEval* (Clark et al., 2011).

By and large, the concatenation models are able to exploit contextual information: BLEU as well as METEOR scores decrease by statistically significant amounts if the context is inconsistent or absent. However, it is difficult to distinguish a winning configuration. In particular, the system that obtains the highest absolute scores is not necessarily the one that learns most from contextual information. The *1Prev+Curr → 1Prev+Curr* system obtains the highest absolute scores among sliding window systems in all four tasks, but is not particularly affected by context inconsistencies. On the other hand, the system using target-language data is most perturbed when context is inconsistent or absent, at least for the OST dataset.[11] It seems therefore that target-language context is as least as important as source-language context. Comparative numbers on the WMT dataset are all very similar, making it hard to draw conclusions.

The 100T fixed-window models perform competitively in terms of absolute scores, compared to the sliding window approaches, despite the alignment problems mentioned above.[12] The compar-

[10] For example, the *1Prev + Curr → Curr* system sees each source sentence twice as often as the *Curr → Curr* system, which might affect general model performance without necessarily improving context awareness.

[11] Note however that we feed the reference instead of the system output at test time for efficiency reasons. Therefore, the numbers cannot be directly compared directly with the other systems, which do not have access to this oracle-type information.

[12] Due to realignment, the number of sentences in the test set varies slightly, which prevents us from computing significance scores. Therefore, the absence of the significance marker * on the *100T → 100T* result lines does not mean that

Dataset:	OST EN → DE						WMT EN → DE					
Context:	Consistent		Incons. (Δ)		None (Δ)		Consistent		Incons. (Δ)		None (Δ)	
System	B	M	B	M	B	M	B	M	B	M	B	M
Curr → Curr (baseline)	21.7	42.6	0.0	0.0	0.0	0.0	39.3	56.9	0.0	0.0	0.0	0.0
1Prev+Curr → Curr	20.9	41.6	-0.3*	-0.5*	-0.2	-0.2	37.6	55.3	-0.5*	-0.3*	-0.2	-0.4*
1Prev+Curr+1Next → Curr	20.1	40.8	-1.0*	-1.2*	-0.6*	-0.5*	34.7	52.3	-0.4*	-0.4*	-0.5*	-0.4*
2Prev+Curr → Curr	20.3	40.4	-0.6*	-0.8*	-0.8*	-0.4*	34.9	53.1	-0.3*	-0.3*	-0.4*	-0.4*
1Prev+Curr → 1Prev+Curr	22.5	43.2	-0.7*	-0.7*	-0.3*	-0.5*	**39.6**	**57.3**	-0.5*	-0.4*	-0.2	-0.3*
1Prev+Curr+1Next → 1Prev+Curr	21.5	42.8	-0.5*	-1.0*	-0.1	-0.6*	38.5	56.0	**-0.8***	**-0.6***	-0.6*	-0.6*
1PrevTarget+Curr → Curr	22.0	42.5	-1.4*	-1.5*	**-1.3***	-1.3*	37.7	55.6	-0.4*	-0.3*	**-0.7***	**-0.7***
100T → 100T	**22.9**	**44.4**	**-1.9**	**-1.9**	-0.5	**-1.8**	39.0	57.2	-0.4	-0.5	0.0	**-0.7**
Selective attention	–	–	–	–	–	–	34.8	53.0	0.0	0.0	-0.2	-0.2

Table 4: BLEU (B) and METEOR (M) scores for EN → DE translation. Absolute scores are reported for the Consistent setting, whereas differences (relative to Consistent) are reported for the Inconsistent and None settings. Statistical significance at $p < 0.05$, obtained by bootstrap resampling, is marked with *.

Dataset:	OST DE → EN						WMT DE → EN					
Context:	Consistent		Incons. (Δ)		None (Δ)		Consistent		Incons. (Δ)		None (Δ)	
System	B	M	B	M	B	M	B	M	B	M	B	M
Curr → Curr (baseline)	27.4	27.6	0.0	0.0	0.0	0.0	34.9	**34.9**	0.0	0.0	0.0	0.0
1Prev+Curr → Curr	26.7	26.8	-0.4*	-0.3*	-0.3*	-0.1*	31.6	32.3	-0.3	0.0	-0.8*	-0.5*
1Prev+Curr+1Next → Curr	24.7	25.5	-0.1	-0.1	-0.3*	0.0	23.0	26.5	-0.1	0.0	**-2.2***	-0.3*
2Prev+Curr → Curr	26.0	26.3	-0.7*	-0.3*	-0.6*	-0.1*	22.0	26.1	-0.1	0.0	-1.3*	**-0.8***
1Prev+Curr → 1Prev+Curr	27.5	27.7	-0.3*	-0.2*	-0.4*	-0.2*	**35.0**	**34.9**	**-0.4***	0.0	-0.9*	-0.5*
1Prev+Curr+1Next → 1Prev+Curr	20.7	24.3	-0.1	0.0	+3.3*	+0.6*	31.2	32.4	-0.3*	**-0.2***	-1.5*	-0.6*
1PrevTarget+Curr → Curr	26.9	27.0	-1.0*	-0.7*	-1.0*	-0.6*	32.7	33.2	-0.3	0.0	-1.1*	-0.5*
100T → 100T	**29.3**	**28.8**	**-1.6**	**-1.0**	**-2.2**	**-1.3**	34.7	**34.9**	+0.1	+0.1	-0.7	-0.3

Table 5: BLEU (B) and METEOR (M) scores for DE → EN translation.

	anaphoric								event	pleonastic	
	it				they		it/they		it	it	
	intra		inter		intra	inter	sing.	group			
	subj.	non-subj.	subj.	non-subj.							**Total**
Examples:	*25*	*25*	*25*	*25*	*10*	*10*	*5*	*15*	*30*	*30*	***200***
OST Curr → Curr	9	7	6	7	5	3	1	5	20	28	91
OST 1Prev + Curr → 1Prev + Curr	10	6	12	9	5	6	1	2	24	25	100
WMT Curr → Curr	14	12	9	10	5	4	0	8	20	26	108
WMT 1Prev + Curr → 1Prev + Curr	9	11	13	12	5	5	1	5	19	28	108

Table 6: Absolute numbers of PROTEST EN → DE pronoun translations evaluated semi-automatically as correct.

DE Pronouns:	ich, es, das, wir, sich, Sie, er, du, sie, die, was, mir, mich, uns, der, man, dich, ihn, dir, dies, ihm, ihr, wer, 's, Ihnen, dem, denen, euch, ihnen, den, Ihr, diese, dessen, deren, einen, dieser, wen, welche, einem, wem, dieses, jene, diesen, dasselbe, welches, einander
Ambiguous:	Sie, den, denen, der, die, diese, dieser, ihm, ihn, ihnen, ihr, man, mich, mir, sich, sie, uns
EN Pronouns:	I, you, it, we, he, what, me, they, who, she, him, them, us, her, himself, itself, themselves, one, yourself, myself, whom, ourselves, i, 'em, herself, mine, yours, ya
Ambiguous:	her, him, it, me, myself, one, she, them, they, us, who, whom, you, yourself
EN Connectives:	although, even though, since, though, meanwhile, while, yet, however
DE Negations:	nicht, nie, niemand, nichts, nirgends, nirgendwo, kein, weder
EN Negations:	no, not, never, nobody, noone, no-one, nothing, nowhere, none, neither, nor

Table 7: List of words and lemmas used to detect discourse-level properties.

ison between consistent, inconsistent and absent context reveals a clear difference between the two datasets: For WMT, the results are almost the same for the three scenarios. This can be attributed to the longer sentences in the WMT test set, which makes the 100 token window performing similar to the one without extended context, as discussed in section 4. In contrast, for the subtitle data, we see notable performance drops when disturbing the model with random or absent context. In this dataset, segments are shorter and 100-token windows substantially increase the context that is available for translation (there are 9.68 sentences per chunk on average).

The selective attention model yields absolute scores with consistent context that are not competitive and barely beat the baseline. It also seems to fail to pick up relevant information from the wider document context, as it obtains almost identical results with inconsistent and absent context.

The WMT EN → DE models have seen back-translations during training but the DE → EN models have not. The results suggest that the additional data helps the models distinguishing consistent from inconsistent input, but further tests will be required to corroborate this hypothesis.

The WMT dataset has shorter documents and longer sentences with more complex discourse-level features. Although this may indicate that it is a more challenging dataset for our models, the performances seem very similar across systems, and it is hard to discriminate informative patterns. However, the inconsistent setting appears to be affected by genre, with none or very small differences with the WMT data, suggesting that the longer sentences are more self-contained in terms of discourse features and that systems effectively pick this signal up. In this same sense (and counterintuitively), the differences between inconsistent and none seem to suggest that as long as the system has access to big enough window, the order in which the document is fed is less important.

5.2 Test suite metrics

Discourse-specific metrics such as Guzmán et al. (2014) would be welcome to assess the translation quality on specific discourse-level features such as those discussed in Section 3.1. However, they have the disadvantage of relying on a discourse parser, which we do not have for German. At least, we are able to evaluate the quality of pronoun translation thanks to the existence of two test suites for English–German pronoun translation: **PROTEST** (Guillou and Hardmeier, 2016; Guillou et al., 2018) is based on TED talks transcripts. These consist of planned speech documents, therefore the genre is somewhere in the middle between news text and dialog. **ContraPro** (Müller et al., 2018) uses material from OpenSubtitles. Due to the overlap of the ContraPro data and our OST training set, we do not use this test suite.

Table 6 reports PROTEST results for two selected systems, the *Curr* → *Curr* baseline and the best-performing variable-window concatenation model *1Prev+Curr* → *1Prev+Curr*. The results draw on a semi-automatic evaluation scheme, where pronouns are accepted as correct if they match the reference and the remaining pronouns are evaluated by hand. The manual evaluation was done by one of the authors.[13] Overall recall of all systems is around 50%, and the differences between systems are quite small.

It can be seen that the models trained on the news dataset obtain higher recall. This confirms our observation in Section 3.1 that the WMT dataset contains higher numbers of coreference chains and cross-sentence pronominal coreference. The context-aware models show small improvements only in the OST dataset. Crucially, the context-aware models show consistently higher numbers in the category of inter-sentential anaphoric pronouns, one of the categories where the previous sentence context is indeed expected to help most. However, most observed differences may not be statistically significant.

The PROTEST evaluation confirms the findings of the WMT18 evaluation (Guillou et al., 2018). In both of these evaluations the *pleonastic* and *event* categories are the least problematic. *Intra-* and *inter-sentential* pronouns are somewhat in the middle but remain difficult, while cases where the anaphor and the antecedent mismatch in features (*they-singular, it/they group*) are very poorly handled.

6 Conclusion

We have presented two English–German document-level translation datasets and shown that they represent different text genres with

the differences are not significant.

[13] We used the provided tool described in Hardmeier and Guillou (2016).

different distributions of discourse-level features. The context-aware NMT models on these datasets show performance differences that are to some extent indicative of the underlying textual characteristics: the longer sentences in the news dataset make it harder to find differences between training configurations or evaluation setups. Fixed-window approaches show surprisingly good results on the movie subtitles dataset, but the impact of the realignment process remains to be investigated further.

The general performance of a document-level MT system can be assessed by testing translation quality with consistent and artificially scrambled context. Models that are able to learn relevant discourse features will be affected if the context is incoherent or absent. Our results show that this test provides a complementary view on the systems' performances.

Our study further suggests that the connections between discourse features and MT results should be analyzed more thoroughly. The detailed breakdown of the distribution of discourse-level properties could be a first step towards the compilation of property-specific test sets.

Automatic measures can be complemented with manual assessment of the outcome from the different test scenarios, which further reveals the effect of discourse features available to the system. We show that pronoun test suites such as PROTEST are a good start for this assessment, although multilingual coverage remains a problem for a systematic evaluation of this kind.

Acknowledgements

The work in this paper was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113), and the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 780069.

The authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussá, 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. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*, pages 128–188, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, 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, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: Induction, search and model errors. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 21–29, Denver, Colorado, USA. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012. N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2012, pages 276–285, Jeju Island, Korea. Association for Computational Linguistics.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

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)*. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, University of the Aegean, Samos, Greece.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, IWSLT 2010, pages 283–289, Paris, France.

Christian Hardmeier and Liane Guillou. 2016. A graphical pronoun analysis tool for the PROTEST pronoun evaluation test suite. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 318–330.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

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

Marcin Junczys-Dowmunt. 2019. Microsoft Translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 424–432, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. Cross-lingual incongruences in the annotation of coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 26–34, Minneapolis, USA. Association for Computational Linguistics.

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, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

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, 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, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Workshop on Hybrid Approaches to Machine Translation at EACL 2012*, HyTra, pages 129—138, Avignon, France.

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

Lesly Miculicich, 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*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, 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: Research Papers*, pages 61–72, Belgium, Brussels. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Kishore Papineni, 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations.

Anita Ramm and Alexander Fraser. 2016. Modeling verbal inflection for English to German SMT. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 21–31, Berlin, Germany. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

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 Zürich.

Daniel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. 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*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.