Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments

Dario Stojanovski **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
{stojanovski,fraser}@cis.lmu.de

Abstract

Cross-sentence context can provide valuable information in Machine Translation and is critical for translation of anaphoric pronouns and for providing consistent translations. In this paper, we devise simple oracle experiments targeting coreference and coherence. Oracles are an easy way to evaluate the effect of different discourse-level phenomena in NMT using BLEU and eliminate the necessity to manually define challenge sets for this purpose. We propose two context-aware NMT models and compare them against models working on a concatenation of consecutive sentences. Concatenation models perform better, but are computationally expensive. We show that NMT models taking advantage of context oracle signals can achieve considerable gains in BLEU, of up to 7.02 BLEU for coreference and 1.89 BLEU for coherence on subtitles translation. Access to strong signals allows us to make clear comparisons between context-aware models.

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015) is a state-of-the-art approach to MT. Standard NMT models translate an input language sentence to an output language sentence, and do not take into account discourse-level phenomena. Cross-sentence context has already proven useful for language modeling (Ji et al., 2015; Wang and Cho, 2016) and dialogue systems (Serban et al., 2016). It has also been of interest in Statistical Machine Translation (SMT) research (Hardmeier, 2012; Hardmeier et al., 2013; Carpuat and Simard, 2012), and NMT research (Wang et al., 2017; Jean et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Tu et al., 2017; Voita et al., 2018).

Two important discourse phenomena for MT are coreference and coherence. Pronominal coreference relates to the issue of translating anaphoric pronouns and is tackled in several works (Guillou, 2016; Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010) and is the central motivation for the DiscoMT shared task on cross-lingual pronoun prediction (Loáiciga et al., 2017). Coherence on the other hand, is important for producing consistent and coherent translations throughout a document, especially for domain-specific terminology (Carpuat, 2009; Ture et al., 2012; Gonzales et al., 2017) and it is helpful to properly disambiguate polysemous words. Modeling discourse-level phenomena for MT is a challenging endeavor because of difficulties in acquiring relevant linguistic signals. Measuring the effect of discourse-level phenomena with automatic metrics such as BLEU is also difficult as pointed out by Hardmeier (2012).

In this paper, we address these issues by proposing several oracle experimental setups for evaluating the effect of coreference resolution (CR) and coherence in MT. Oracle experiments provide strong linguistic signals that enable strongly visible effects on BLEU scores, thus alleviating the difficulty of using BLEU to evaluate discourse-level phenomena in MT. Oracles highlight the capability of NMT systems to use context (which we call context-aware NMT) and to handle different discourse-level phenomena. They provide a variety of scenarios that can easily be set up for any domain, dataset or language pair, unlike discourse-specific challenge sets (Bawden et al., 2018) which must be manually created. Furthermore, strong linguistic signals from oracles enable us to easily study how the models use context.

Our primary task is translating subtitles from English to German. Subtitles provide for a reasonable diversity of topics necessary for testing coherence. They also contain a large amount of short, informal and conversational text, where anaphoric pronouns are very important. We study coreference by aiding pronoun translation and coherence

Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 49–60

https://doi.org/10.18653/v1/W18-6306

by providing disambiguation signals for translation of polysemous words. The oracles are automatically created and targeted for each discourse phenomenon. We additionally include a previous target sentence oracle, where the context consists of the previous target sentence, as a more generic way of including context. This is an interesting oracle, but this scenario is actually also beneficial for online post-editing, because the gold standard previous target sentence is available there.

We propose a simple, yet effective extension to standard RNN models for NMT (which we refer to as NMT(RNN)) which models context by employing attention over word embeddings only. We compare it against a standard NMT(RNN) model working on a concatenation of consecutive sentences (Tiedemann and Scherrer, 2017). Additionally, we evaluate the Transformer (Vaswani et al., 2017) and propose a context-aware NMT(Transformer) extension. Our oracles allow us to compare the context-aware NMT models with the baselines and make strong conclusions. Moreover, we study how comparable oracles are with the challenge sets proposed by Bawden et al. (2018) by analyzing the performance of our context-aware model with both approaches. Finally, we conduct a qualitative study and show the inner workings of context-aware models under different oracle settings.

Contributions: (i) We modify the data using an oracle experimental setup in order to accommodate evaluating coreference and coherence in NMT. (ii) Our evaluation is independent of carefully constructed challenge sets, and can easily be transferred across language pairs and domains. (iii) Results clearly show context-aware NMT(RNN) and NMT(Transformer) can improve performance over NMT models without access to context. (iv) We empirically analyze the pros and cons of the major approaches to context-aware NMT and explain how different modeling decisions interact with different discourse phenomena. (v) We present the trade-offs in modeling power versus speed that are important when considering multiple sentences of context.

2 Oracle Signals for Coreference and Coherence

Acquiring clean and strong context signals is a difficult challenge and previous work has not proposed a way to do this on a larger scale. In our work, we use oracles, where the context signals are strong and allow us to carry out clear analysis. We define three oracles which differ based on the context supplied to the model.

First, we define the previous target sentence oracle where the context is the gold standard previous target sentence. Second, we define the coreference or pronoun oracle where we simulate perfect knowledge of gender and number for pronoun translation. Finally, we define the coherence or more specifically, the repeated words oracle where we help in identifying polysemous words and providing the correct signal for disambiguation.

Each of these oracles is accompanied by a fair and a noisy oracle experimental setup. For the fair setup, we obtain the linguistic signals in a realistic way without having access to any target side knowledge. In the noisy oracle setups, we add additional target side information to the oracle signals. This additional information is not necessarily relevant to the specific problem at hand (coreference or coherence) and it is used to test the robustness of the models to identify the proper signals.

The oracle datasets are created in an automatic way. We only need to manually define the list of pronouns that will be taken into consideration in the coreference oracle.

Oracle Table 1 shows samples from our oracle setup. For each example we show the context, original source sentence, our modified oracle sentence and the target sentence. The first two examples show coreference (pronoun) oracle samples, while the third one a coherence (repeated words) oracle sample. The text in brackets shows which is the counterpart repeated target word or the gender of the noun the pronoun is referencing. It is not explicitly provided to the models. The text preceding the special token *!@#$* in the oracle examples is the input to the context part of the architecture.

For coreference, we aid the model with pronoun translation as can be seen in example (c). In this case, *it* refers to *Roman* (meaning *novel*), which is apparent in the previous sentence (a). Without this information the model will have difficulties generating the proper translation *er* (the German masculine pronoun agreeing with *Roman*).

When creating the pronoun oracle setup, we do not utilize the context sentence. Instead, we just consider the current source and corresponding target sentence. If both sentences contain at least one pronoun in their respective languages, we mark

the source pronouns with XPRONOUN and insert the target pronouns in the context of the main sentence, as in example (c).

The example shows that the context provides access to perfect knowledge of the coreferent, which in turn tells us the number and gender. However, the models still need to learn to use the correct pronouns. As we can see in example (g), there may be multiple pronouns in the context. Since (g) is an imperative sentence, *Sie* does not have a pronoun counterpart in the source and it is used in conjunction with the German verb for *use*.

Example (k) shows how we model the coherence phenomenon by using repeated words. Given the English word *source* in a sentence without helpful context, it would be impossible to disambiguate between two possible translations of the word: *Quelle* (a source of a fountain or figuratively the source of information) or *Ursprung* (origin, where something originates from). However, we see that the previous sentence (i) contains the relevant information to select the correct translation of the English *source*. The word *source* is present in the previous and current source sentence and *Ursprung* is present in the previous and current target sentence. When we find at least one repeated word on both the source and target side, we mark the source word with a special token XREP and the repeated target word is used as context to the main source sentence. The intuition here follows previous work (Tu et al., 2017) where past translation decisions are used for disambiguation. This oracle is admittedly weaker than the coreference one since it relies on the assumption that a polysemous word has already been seen in the text. However, if a word occurs in two consecutive sentences, it is likely that it will have the same translation.

For the previous target sentence oracle, we use the gold standard previous target sentence as context and don't modify the main source sentence. We also setup experiments with 2 and 3 previous target sentences as context.

Fair For the fair coreference setup, we attempt to acquire gender and number knowledge by using a coreference resolution tool, namely CorefAnnotator from Stanford CoreNLP[1] (Clark and Manning, 2016a,b). We run the model on entire documents. We only modified sentences that contain a pronoun which has an antecedent in the previous source sentence. Consequently, the pronoun is marked and the antecedent is inserted into the context of the given sentence. In this way, we don't utilize any target side knowledge.

[1] https://stanfordnlp.github.io/CoreNLP

context sentence
(a) Let me summarize the novel[masculine] for you.
source sentence
(b) It presents a problem
pronoun oracle sample
(c) er[masculine] !@#$ XPRONOUN It presents a problem.
target sentence
(d) Er präsentiert ein Problem.
context sentence
(e) But you have a charm[masculine] everyone else here seems to respond to.
source sentence
(f) Use it. OK, sport?
multiple pronoun oracle sample
(g) Sie ihn[masculine] !@#$ Use XPRONOUN it. OK, sport?
target sentence
(h) Setzen Sie ihn ein.
context sentence
(i) When dealing with a crisis everyone knows you go right to the source[Ursprung].
source sentence
(j) God the source is pretty.
repeated words oracle sample
(k) Ursprung !@#$ God the XREP source is pretty.
target sentence
(l) Mann, so ein hübscher Ursprung.

Table 1: Coreference and coherence oracle samples. For detailed explanation of the examples, refer to Section 2.

For the fair coherence experiment, we don't have access to target side information and we just put special emphasis on words that are polysemous candidates. As a result, we only use repeated source words. A repeated word is marked in the main sentence and it is used as context.

For the fair previous sentence experimental setup, we use the same models trained on the previous target sentence oracle setup, but evaluate them by translating the previous source sentence with a baseline model and using this translation as context. Additionally, we train models where the previous sentence is from the source side.

Noisy oracles In order to test the robustness of context-aware models, we define noisy coreference oracles. We use the same approach as in the oracle, but the previous gold standard target sentence is added at the beginning of the context (which already contains the target side pronouns).

We also define noisy oracles for coherence. In this case, this is achieved by marking repeated source words and marking repeated target words in the previous target sentence and using the modified previous target sentence as context.

3 Related Work

Bawden et al. (2018) is a recent work with similarities to ours. They look at the scores computed by context-aware models using challenge sets, by comparing model scores on two perfect target language sentences differing only on a single choice of, e.g., gender for a pronoun, and providing two different contexts to try to obtain, e.g., masculine in the first case and feminine in the second case.

Like Bawden et al. (2018), we provide a focused evaluation on coherence and coreference, but unlike their work, we do not depend on manually created datasets. Our simple oracles are a strong alternative to manually constructed challenge sets, as we can easily have a more diverse experimental setup (our oracles can be defined for different languages, domains and datasets with little effort).

Several approaches have been proposed for context-aware NMT that utilize a separate mechanism to handle extra-sentential information. Wang et al. (2017) integrate cross-sentence context using gates in the decoder, which control information flow between the cross-sentence context and the current decoder state. However, the context representation is fixed at each decoding time step, while the model needs to focus on different parts of the context. Tu et al. (2017) propose a caching mechanism that stores previous translation decisions. As a result, this approach fails to take into account CR as stored translation decisions can't be used to address this phenomenon. Jean et al. (2017) and Bawden et al. (2018) propose methods using a separate RNN-based context encoder. Tiedemann and Scherrer (2017), propose concatenating the preceding sentence, both on source and target side and then using a standard NMT model. These approaches are computationally expensive. They either have an extra RNN-based encoder (Jean et al., 2017; Bawden et al., 2018) or work on very long sentences (Tiedemann and Scherrer, 2017).

A recent work by Voita et al. (2018) proposed a context-aware Transformer model and provided an analysis of anaphora resolution in MT. Their proposed model is conceptually similar to our NMT(Transformer) model, differing in that the context is integrated in the encoder unlike our model which does it in the decoder.

We propose a simple NMT(RNN) model that only uses attention to encode the context and integrates it with a gating mechanism (Wang et al., 2017). It provides for a better computational efficiency compared to models employing an extra RNN-based encoder. We also propose a context-aware Transformer model. In the experiments, we compare our models against a concatenation NMT(RNN) and NMT(Transformer) model (Tiedemann and Scherrer, 2017).

4 Context-Aware Models

4.1 Lightweight context-aware NMT(RNN) model

In this paper, we introduce a new lightweight context-aware model based on the attention encoder-decoder model proposed by Bahdanau et al. (2015). We introduce this context-aware model to compare against the proposed model by Tiedemann and Scherrer (2017) as an alternative approach to handling context.

The encoder part of the model, takes the source sentence $X = (x_1, x_2, \ldots, x_{T_x})$ and generates a set of annotation vectors $\{h_1, h_2, \ldots, h_{T_x}\}$ where $h_i = \left[\overrightarrow{h}_i; \overleftarrow{h}_i\right]$. $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ are the i-th hidden states from the forward and backward recurrent networks respectively. The decoder generates one target symbol y_i at a time by computing the conditional probability $p(y_i|y_1, y_2, \ldots, y_{i-1}, x) = f(y_{i-1}, s_i, c_i)$ where c_i represents the attention weighted sum of annotation vectors and is computed as in (Bahdanau et al., 2015). Unlike previous approaches that model context by employing an RNN-based encoder (Jean et al., 2017; Bawden et al., 2018), we propose to utilize the capability of the attention mechanism only. This provides for better computational efficiency, thus allowing the model to exploit larger context at a lower computational cost.

The context sentence is given as a sequence of $X^c = (x_1^c, x_2^c, \ldots, x_{T_x^c}^c)$. We map the tokens to the corresponding word embeddings w_i^c. We share all embeddings across the model, including the context ones. The attention on the cross-sentence context is conditioned on the previously generated token y_{i-1} current candidate decoder state s_{i-1} and attention weighted main sentence representation c_i. Formally, the context sentence representation is computed as $c_i^c = \sum_{j=1}^{T_x^c} \beta_{ij} w_j$ where $\beta \propto exp(f_{att}^c(y_{i-1}, s_{i-1}, w_j, c_i))$.

We integrate the context representation using a gating mechanism (Wang et al., 2017) which controls the flow of information between the current decoder state and the context representation. which is computed as $g = f_g(y_{i-1}, s_{i-1}, c_i, c_i^c)$.

The final decoder representation is computed as $s_i = f_c(y_{i-1}, s_{i-1}, c_i, g \otimes c_i^c)$.

4.2 Transformer context-aware model

The Transformer (Vaswani et al., 2017) is an encoder-decoder architecture which fully relies on attention. The encoder layers have two main components, a multi-head self-attention and a position-wise fully-connected feed-forward network. Each of these components is followed by a residual connection. In the self-attention sublayer, each word from the input sentence acts as a query, key and value when computing the attention. Each attention head uses the queries and keys to compute a dot product to which a softmax is applied in order to get the attention weights to score the values. Consequently, the representation of each word depends on all the others. The final representation is generated by concatenating the output of the separate attention heads and inputting it to the feed-forward network. The decoder on the other hand, has three sublayers. It starts by applying masked self-attention which is then used to compute multi-head attention over the encoder representation. This is then used as input to a feed-forward network as in the encoder.

The proposed context-aware model in this paper is built as an extension to the standard Transformer. All embeddings including the context embeddings are shared across the model. We modify the encoder by sharing the parameters for the multi-head self-attention for the main and context sentence. However, we don't share the feed-forward network after the self-attention.

The standard decoder computes a multi-head attention c_i over the main encoder representation using the output from the masked self-attention c_i^m. We add an additional multi-head attention over the context representation c_i^c as well. Before computing the context attention, the output of the masked self-attention is projected using a feed-forward network. The main and context multi-head self-attention representations are merged using a gating mechanism as $s_i = g_i \otimes c_i + (1 - g_i) \otimes c_i^c$ where $g_i = \sigma(W_e c_i + W_c c_i^c + W_m c_i^m)$.

5 Experiments

We train our models on OpenSubtitles2016 En-De with $\approx$ 13.9M parallel sentences. The development and test set consist of 6 and 7 documents randomly sampled from the dataset, containing 3172 and 4627 sentences respectively. In the coreference oracle setup $\approx$ 7.8M training samples were modified and added the appropriate context, while in the coherence setup only $\approx$ 0.8M. The remaining samples are unchanged and have no context.

We apply tokenization, truecasing and BPE splitting computed jointly on both languages with 59500 operations. All sentences with length above 60 tokens are discarded. Batch size is 80. All embeddings are tied (Press and Wolf, 2017) including the ones in the context part of the architecture. Dropout (Gal and Ghahramani, 2016) of 0.2 is applied and 0.1 on the embeddings. We apply layer (Ba et al., 2016) and weight normalization (Salimans and Kingma, 2016). The models are trained with early-stopping based on the development set's cost. We report BLEU score on detokenized text.

Our RNN-based model is implemented as an extension to Nematus[2] (Sennrich et al., 2017). We used the Sockeye[3] (Hieber et al., 2017) implementation of the Transformer. For the Transformer we use hyper-parameters as similar as possible to the ones in the Nematus models. We additionally use label smoothing of value 0.1. Both, the baseline and context-aware model have 4 layers. We didn't do any special hyper-parameter tuning for the context-aware models, so further performance improvements are possible. The datasets and the source code for our context-aware models are publicly available[4].

6 Experimental Results

6.1 Previous target sentence oracle

In this section, we discuss the effect of using context in context-aware NMT. In Table 2 we show the results for the three different oracle setups. Experiment (1a) shows that a baseline NMT(RNN) model obtains 28.57 BLEU on the test set. The NMT(Transformer) baseline (1b) on the other hand, achieves 29.53 BLEU. Using the gold standard previous target sentence as context, provides for 1.32 BLEU improvement on the test for our context-aware NMT(RNN) model (2a) and 1.78 BLEU for the concatenation NMT(RNN) model (3a). Our proposed context-

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

[3] https://github.com/awslabs/sockeye

[4] http://www.cis.uni-muenchen.de/~dario/projects/oracles

aware NMT(Transformer) model (2b) also improves upon the baseline, but only by 0.6 BLEU, and the concatenation model (3b) closely follows the RNN model, adding 1.49 BLEU.

We also evaluate the usefulness of larger context. Using the previous 2 (6a) and 3 (7a) sentences consistently adds $\approx$ 0.6 BLEU with the concatenation NMT(RNN) model. The context-aware NMT(RNN) model, does not improve when using 2 sentences (4a), but has large gains when extending to 3 (5a). In our context-aware models, the larger context is handled by concatenating all previous sentences. The context-aware NMT(Transformer) (4b), (5b) was actually hurt by the larger context. On the other hand, for the concatenation model (6b), (7b) we observed some improvements, but they were not as consistent as the gains for the NMT(RNN) model.

The results in (2ab), (3ab), (4ab), (5ab) (6ab), (7ab) are obtained with models trained and evaluated with the gold standard previous target sentences as context. In the fair experiments (8ab), (9ab) we train with the gold standard previous target sentence as context, but then evaluate with translations of the previous source sentences obtained with the baseline model. This lowers the performance of both NMT(RNN) models (8a), (9a), but they still improve over the baseline. Our context-aware NMT(Transformer) model (8b) slightly lowers performance compared to the baseline, unlike the concatenation model (9b).

Additionally, we train context-aware models where the previous sentence is obtained from the source side (10ab), (11ab). Even in such a scenario, context-aware and concatenation NMT(RNN) models obtain improvements over the baseline. Again, the concatenation NMT(Transformer) shows improvements over the baseline. The context-aware NMT(Transformer) was not able to make use of the source side information. Given that the encoder representations are shared this is to some extent surprising and suggests that additional encoder components are necessary to model the contextual representation.

6.2 Coreference

Results for coreference are also shown in Table 2. Experiments (12a) and (12b) show the results we obtained with the pronoun oracle setup. It is clear that NMT can benefit from strong coreference signals. We observed a large difference between the

	(a) RNN	(b) TF
(1) baseline	28.57	29.53
(2) context - gold prev. target	29.89	30.13
(3) concat - gold prev. target	30.35	31.02
(4) context - gold prev. 2 target	29.96	29.57
(5) context - gold prev. 3 target	30.95	29.98
(6) concat - gold prev. 2 target	30.96	31.69
(7) concat - gold prev. 3 target	31.56	31.26
(8) context - baseline prev. target	29.10	29.25
(9) concat - baseline prev. target	29.28	29.89
(10) context - prev. source	29.48	28.80
(11) concat - prev. source	29.56	30.25
Coreference		
(12) context - pronoun oracle	34.35	34.60
(13) context - fair	29.05	28.76
(14) context - noisy pronoun oracle	33.61	34.62
(15) concat - noisy pronoun oracle	35.59	35.18
Coherence		
(16) context - repeated target words	29.83	29.35
(17) context - repeated source words	29.27	29.04
(18) context - noisy rep. target words	30.07	29.85
(19) concat - noisy rep. target words	30.46	31.25

Table 2: BLEU scores from all of the oracle experimental setups on the test set. Results in the first column correspond to the NMT(RNN) context-aware and concatenation models while the second column to the NMT(Transformer) ones. The number in brackets in each line is used to indicate the corresponding experiment throughout the text.

improvements on the development and the test set, probably because this phenomenon is not equally prominent in the datasets. In the absence of perfect CR, this setup is a reasonable proxy for obtaining coreference signals and gender information, and the context-aware models achieve large improvements over their respective baselines.

Experiments (13a) and (13b) show the results for the fair coreference setup. Using a CR tool, we identified the appropriate antecedents (to current sentence pronouns) in the previous source sentence and used them as context. The results show small improvements on the test set. This signal is significantly weaker. Moreover, only $\approx$ 0.3M samples had a non-empty context, meaning a pronoun was referring to a coreferent as identified by the CR tool. These results show that while weak, the context-aware NMT(RNN) model is able to utilize this signal. The NMT(Transformer) model on the other hand, was significantly hurt by this setup. We attribute this to the model not being able to handle scenarios where the majority of the samples are without context information.

In the noisy pronoun oracle setup, the context consists of the previous gold standard target sentence to which we append the target side pronouns as in the previously outlined pronoun oracle setup. The results are shown in Table 2. We can ob-

serve that the context-aware NMT(RNN) model (14a) is actually hurt by the extra information in the form of previous target sentence. We attribute the decrease to the model learning to strongly attend to all pronouns in the context. As such, in some cases, it chooses to attend to a pronoun from the previous sentence which ends up acting as noise in these models. Using oracles allowed us to easily find this important weakness in our model design. The context-aware NMT(Transformer) model (14b) is more robust to noise and had no problems identifying the appropriate information.

Using the same setting for the concatenation NMT(RNN) model (15a), achieves best performance with an absolute gain of 7.02 BLEU. Based on the obtained results in (3a), we conclude that the effects in (15a) are a compound of the capability of concatenation models to make use of the previous sentence and target side pronouns. The same effects can be observed for the NMT(Transformer) concatenation model as well (15b). However, despite the concatenation Transformer being able to obtain better results for the previous target sentence and pronoun oracle than the RNN model, the compound effect is not as strong.

6.3 Coherence

Table 2 shows the results we obtained for the coherence experimental setup. For the oracle setup, we identify repeated source and target words in the previous and current sentence, mark the source words and insert the target words in the context. For the fair setup, we insert repeated source words in the context. The aim with this scenario is to emphasize which words are potentially important for disambiguation. Moreover, in the oracle setup, we provide the presumably gold standard translation of the repeated word in the appropriate context.

Both scenarios (16a), (17a) obtain improvements over the baseline with the NMT(RNN) model, although not as strong as the gains with the pronoun oracle. One reason is that the number of samples with context is significantly smaller than the pronoun oracle. Another potential reason is that coherence is already modeled well by the baseline. The results indicate that obtaining coherence and disambiguating signals from past translation decisions, whether from an oracle such as in our work or from the model itself (Tu et al., 2017) is difficult. Nevertheless, the noticeable gains in BLEU we observed in our experiments confirm that further improvements can be made. The context-aware NMT(Transformer) is hurt by these oracle setups as shown in experiments (16b) and (17b) because of the lack of sufficient context.

Table 2 presents the results for the noisy coherence oracle. The context-aware NMT(RNN) model (18a) obtains improvement over the baseline of 1.5 BLEU and the concatenation model (19a) of 1.89 BLEU. This is likely a compound effect of having access to the entire previous target sentence as in (2a) and (3a) and the weak signals in the form of pointers to where disambiguation is necessary. This is to some extent matched by the Transformer experiments (18b), (19b).

6.4 Comparison with challenge sets

In order to assess the quality of our oracles, we also set them up on OpenSubtitles2016 En-Fr and compare them against the challenge sets proposed in Bawden et al. (2018). This allows us to compare the two methods and show whether we can draw similar conclusions about a model when evaluating it with both the oracles and challenge sets. For simplicity, we only evaluate our proposed context-aware NMT(RNN) model. We randomly sampled documents from the En-Fr dataset to create a development and test set. The challenge sets are used as provided by Bawden et al. (2018). We set up the oracles in the same way as for En-De. However, in French the pronouns *le*, *la* and *les* can also be used as definite articles. Therefore, we used MarMoT (Mueller et al., 2013) to filter out these instances.

We compare the methods by measuring the improvements a context-aware model achieves over a baseline, on our oracles and on the challenge sets. Since our oracles use target side knowledge, we use the version of the challenge sets where the previous sentence is from the target side. This provides for a fairer comparison. We train our context-aware model on the pronoun and repeated words oracle. In order to evaluate the model on the challenge sets, we train the model with the gold standard previous target sentence as context.

The baseline model obtains a score of 27.73 BLEU on the test and by design, it achieves 50% accuracy on the coreference and 50% accuracy on the coherence challenge set. Our proposed context-aware model trained on the pronoun oracle achieved 30.72 BLEU on the test set. On the repeated words oracle, it scored 28.25 BLEU. As in the En-De experimental results, our model ob-

pronoun oracle	meine er !@#$ XPRONOUN My reading of the prophecy is that XPRONOUN it will come in 2012
reference	Meine Textstudien ergeben, daß er 2012 kommen wird
baseline	Mein Lesen der Prophezeiung lautet, dass *es* 2012 kommen wird
context	Meine Lesung der Prophezeiung ist, dass er 2012 kommen wird
repeated words oracle	Abneigung Romulaner !@#$ If you had seen them kill your parents, you would XREP understand it is always the XREP time for those XREP feelings.
reference	Höatten Sie mit angesehen, wie Ihre Eltern getötet werden... Meine Abneigung gegen die Romulaner ist universell.
baseline	Wenn du gesehen hättest, wie sie deine Eltern töten würden, würdest du verstehen, dass es immer die Zeit für diese *Gefühle* ist.
context	Wenn du gesehen hättest, wie sie deine Eltern getötet haben, würdest du verstehen, dass es immer die Zeit für diese Abneigung ist.
prev. sent. oracle	Er dachte, die Geschichte handelte von einem Fisch. !@#$ It isn't?
reference	Tut sie nicht?
baseline	Ist *es* nicht?
context	Ist *es* nicht?

Table 3: Samples from the qualitative analysis.

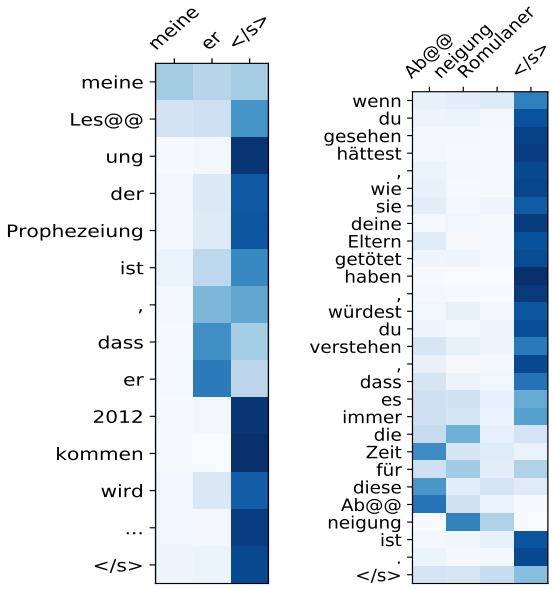
tains small gains for coherence and larger ones for coreference. The context-aware model we trained with the previous target sentence as context, scored 63.0% and 54.0%, on the coreference and coherence challenge set, respectively. From these results we also can conclude that our model is reasonably powerful to handle coreference and marginally improves coherence. These results show that challenge sets and oracles provide comparable results when evaluating discourse in MT. However, our oracle setups are easier to define and control.

6.5 Qualitative study

In this section, we show examples from our oracle setups and provide visualizations of the extra-sentential attention for our context-aware and the concatenation NMT(RNN) model (Tiedemann and Scherrer, 2017). We also show the activations of the decoder gates which control the context information flow. This can help us understand how the models make decisions at each time step.

In Table 3 we show the pronoun, repeated words and previous target sentence oracles and compare the output from a baseline and our proposed context-aware model against the reference translation. For simplicity, in the visualizations for the concatenation model, we only present the attention over the previous sentence and the sentence separating token SEP.

The first row in Table 3 shows a pronoun oracle sample. In this case, *it* refers to *comet*. It is obvious that there is not sufficient information in the main sentence alone to properly translate *it* and the baseline model falls back to the data-driven prior, which is to generate *es*.



(a) Pronoun (b) Repeated words

Figure 1: Context attention for the pronoun and repeated words oracles.

From the visualization in Figure 1a we see that our context-aware model pays attention to the appropriate pronoun (*meine*, *er*). From Figure 3 we see that for this example, the noisy oracle shows the same behavior and correctly ignores the noise. Furthermore, Figure 2a and Figure 2b show that the gate activations follow the intuitive assumption that they should be high when generating pronouns. Our model in the noisy pronoun oracle produced a correct translation, but it still weakly paid attention to irrelevant parts of the sentence. From Figure 4 we see that concatenation model on the other hand, makes a clean distinction between what is relevant and what is not, and only has strong attention over the pronouns.

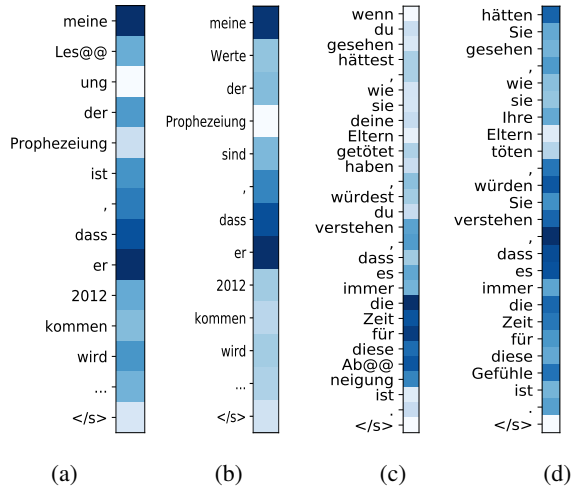


Figure 2: Gate activations for pronoun and repeated words oracles. (a) pronoun oracle, (b) - noisy pronoun oracle, (c) - repeated words oracle, (d) - noisy repeated words oracle.

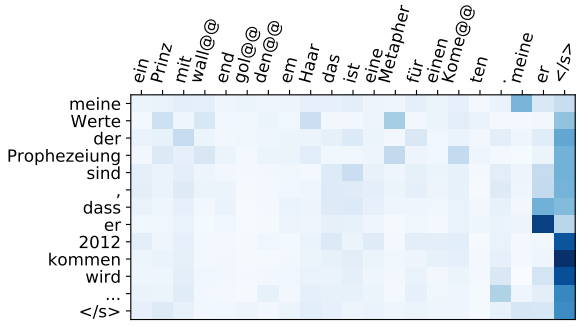


Figure 3: Context attention of our proposed model on the noisy pronoun oracle.

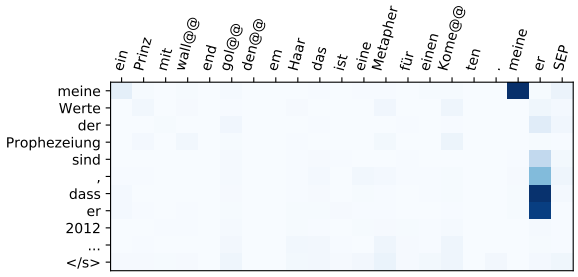


Figure 4: Attention over the previous sentence of the concatenation model on the noisy pronoun oracle.

The second sample is selected from the repeated words oracle setup. Because the reference translation does not exactly match the source sentence, there is a small mismatch between the repeated words on the source and target side. However, we see that without the contextual signal that *feelings* in this case refers to *adverse feelings* (as indicated by *Abneigung*) the baseline falls back to the more common translation *Gefühle*. We also looked at the previous sentence which did not have any context information and both the baseline and the context-aware model generated *Gefühle*.

Figure 1b shows that the context-aware model has no problem attending to the disambiguating signal (*Abneigung*) and it also uses this signal when generating the determiner *dieses* which is dependent on the noun. However, we also can observe that given the incorrect indication to look at the context when translating *time*, it also has attention activation over the context as well. This is closely followed by the gate activations in Figure 2c. The same doesn't happen when translating the marked source token *understand*. This is probably because the model is confident that it doesn't need context when translating *understand*.

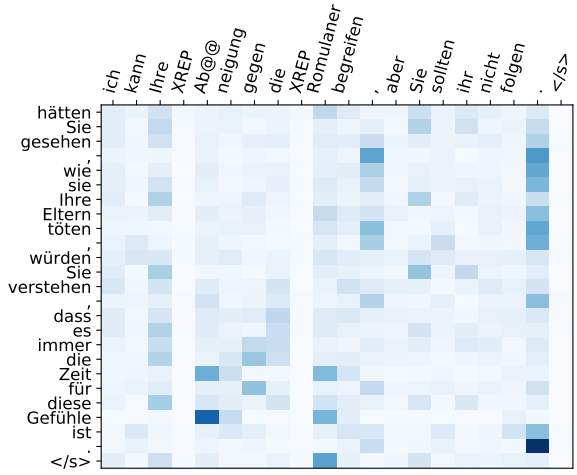


Figure 5: Context attention of our proposed model on the noisy repeated words oracle.

From Figure 5 and Figure 2d we see that the context-aware model in a noisy repeated words oracle setting has difficulties identifying the coherence information and when to use it. It tends to pay attention to certain words throughout the whole sequence generation. This is likely a side effect of having access to the previous target sentence which in other cases provides useful information. Although it pays attention to the appropriate repeated word (*Abneigung*), it still fails to generate it. Since the concatenation model uses an RNN over the context, it has no problem identifying the disambiguating signal, marked with XREP and generates it accordingly (Figure 6).

We also did an analysis of the previous target sentence oracle as well as the models that use the previous source sentence as context. We looked at examples where there is an anaphoric pronoun *it*. When the context is from the source side, our

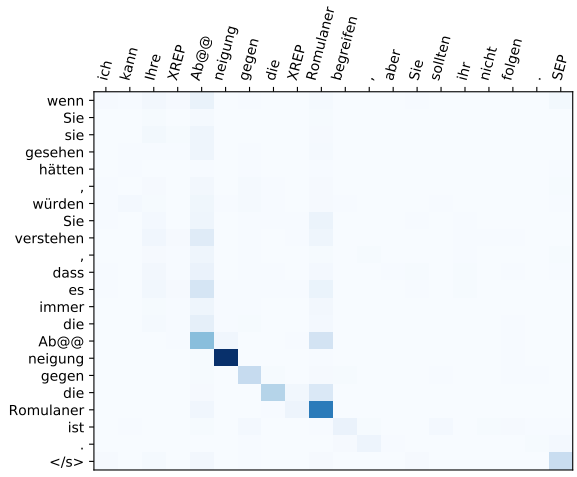


Figure 6: Attention over the previous sentence of the concatenation model on the noisy repeated words oracle.

context-aware model tends to pay attention to a single noun, while in the previous target sentence oracle, it looks at more explicit gender information, such as pronouns, articles etc. This is illustrated in the last example in Table 3 and Figure 7 and 8. In this case, *it* refers to *die Geschichte* or *story*. When translating *it* both models paid attention to the appropriate place in the previous sentence, but failed to generate the correct pronoun *sie*. For this particular example, the concatenation model paid no attention to the previous sentence.

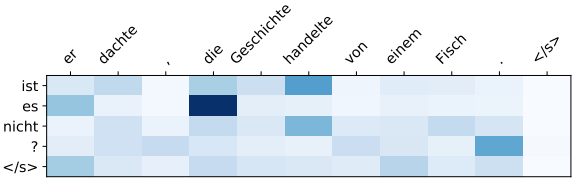


Figure 7: Context attention of our proposed model on the previous target sentence.

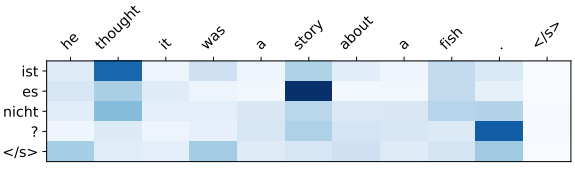


Figure 8: Context attention of our proposed model on the previous source sentence.

6.6 Model inference speed

Although the concatenation model performs better than our context-aware model, an important consideration when working with context-aware NMT is computational efficiency. We compared inference times for the RNN models on the development set. We report times with context size of 1, 2 and 3 previous sentences.

The context model took 1233 seconds to decode the development set, while the concatenation model 2063 seconds. The concatenation model took additional ≈ 900 seconds for each additional context sentence. Because our context-aware implementation is not tightly dependent on context length, there are no considerable drops in speed. This is a disadvantage of the concatenation approach. If one is to use large context, or even entire documents, the problem quickly becomes very computationally expensive. This highlights the necessity of specialized context-aware models. Since the Transformer can be more easily parallelized, there is still room for improving the computational performance of our context-aware Transformer. As a result, we leave such a comparison for future work.

7 Conclusion and Future Work

We used simple oracles to look at discourse-level phenomena in MT. We compared context-aware NMT models and show that these approaches provide large gains in BLEU for coreference and coherence given clear oracle signals. We also showed that even when using fair signals, such as the previous source sentence or a system translation of the previous target sentence, NMT models benefit and make use of the extra information. Some future work in context-aware NMT can focus on using the standard NMT architecture, which performs well. However, if one requires access to larger context, vanilla NMT will have difficulties scaling in terms of speed and perhaps even in modeling ability. For this reason, a promising way forward is studying different ways of modeling and integrating context that support fast inference. Oracle experiments will allow us to quickly test interesting modeling differences.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable input and Daniel Ledda for his help with examples. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15. ArXiv: 1409.0473.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *NAACL 2018*, New Orleans, USA.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27.

Marine Carpuat and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

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.

Liane Kirsten Guillou. 2016. *Incorporating pronoun function into statistical machine translation*. Ph.D. thesis, The University of Edinburgh, UK.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics); 4-9 August 2013; Sofia, Bulgaria*, pages 193–198.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints*.

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

Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document context language models. *arXiv preprint arXiv:1511.03962*.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261.

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 *The Third Workshop on Discourse in Machine Translation*.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163.

Tim Salimans and Diederik P Kingma. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909.

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.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016.

Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

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.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2017. Learning to remember translation history with a continuous cache. *arXiv preprint arXiv:1711.09367*.

Ferhan Ture, Douglas W Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426.

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 *Advances in Neural Information Processing Systems*, pages 6000–6010.

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.

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.

Tian Wang and Kyunghyun Cho. 2016. Larger-context language modelling with recurrent neural network. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 3, pages 1319–1329.