Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning

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

Abstract

Modeling anaphora resolution is critical for proper pronoun translation in neural machine translation. Recently it has been addressed by context-aware models with varying success. In this work, we propose a carefully designed training curriculum that facilitates better anaphora resolution in context-aware NMT. As a baseline, we train context-aware models as was done in previous work. We leverage oracle information specific to anaphora resolution during training. Following the intuition behind curriculum learning, we are able to train context-aware models which are improved with respect to coreference resolution, even though both the baseline and the improved system have access to exactly the same information at test time. We test our approach using two pronoun-specific evaluation metrics for MT.

1 Introduction

Modeling gender-pronoun agreement and anaphora resolution in machine translation is difficult because most models work on individual sentences. In many cases the antecedent noun is not present in the sentence being translated, but is rather in a preceding sentence. Sentence-external anaphora are a problem in many domains (e.g., consider conversational texts). NMT models can be extended to receive the previous sentences of a document as input. Previous context-aware NMT models include (Jean et al., 2017; Wang et al., 2017; Tu et al., 2018; Voita et al., 2018; Stojanovski and Fraser, 2018; Zhang et al., 2018a; Miculicich et al., 2018). Previous work on evaluation has shown that context-aware NMT improves over sentence-level baselines, both in terms of BLEU and in terms of metrics tailored for pronoun evaluation (Bawden et al., 2018; Voita et al., 2018; Müller et al., 2018).

In this work, we propose a technique for improving the ability of context-aware models to handle anaphora resolution. The technique is based on curriculum learning (Bengio et al., 2009) which proposes to train neural networks in a similar fashion to how humans learn. Curriculum learning is a method that proposes training neural networks by gradually feeding increasingly more complex data instead of training models by randomly showing data samples.

We borrow on the intuition behind curriculum learning by initially training models with a form of "training wheels", where the anaphora relationships are made explicit. We take the key idea from previous work, which is to use gold-standard reference pronouns as oracles (Stojanovski and Fraser, 2018). We then gradually remove the oracles in consecutive fine-tuning steps, until we have a model working without oracle information. We expect that explicitly showing the reference pronouns in the context will make it easier to model the gender of antecedent nouns and bias the model to do more aggressive anaphora resolution when encountering ambiguous pronouns in the source language (the translation of ambiguous pronouns depends on the antecedent). We experimentally show the importance of the learning rate when training context-aware models with regards to our curriculum learning approach on both pronoun and overall translation performance. For this

reason we present experiments training context-aware models with low and high initial learning rates. Note that our approach could be extended to other discourse-level phenomena, provided that useful oracles are easily obtainable. Our main contributions are: 1) We propose a curriculum learning method that supplies oracle information in training (but not testing) to improve anaphora resolution in NMT. 2) We show that our method works when training models with a low learning rate according to different metrics (measuring both MT quality overall and pronoun correctness). 3) We outline best practices for training and fine-tuning context-aware models.

2 Related Work

Several works have proposed methods and models of including contextual information (Wang et al., 2017; Jean et al., 2017; Bawden et al., 2018; Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Zhang et al., 2018a; Kuang and Xiong, 2018; Kuang et al., 2018). In general, these models make use of extra-sentential attention conditioned on the main sentence being translated and use gates to control the flow of contextual information. The model we use is based on these general concepts as well.

Improvements in BLEU cannot be conclusively attributed to improved anaphora resolution and therefore additional metrics are required. Several works have proposed methods of evaluation and have shown that context-aware NMT achieves improvements. Müller et al. (2018) propose an automatically created challenge set where a model scores German translations of an English source sentence. The source sentences contain an anaphoric third person singular pronoun and the possible translations differ only in the choice of the pronoun in German. Bawden et al. (2018) is an earlier work proposing a manually created challenge set for English and French. Miculicich et al. (2018) evaluate their model's effectiveness on pronoun translation by computing pronoun accuracy based on alignment of hypothesized translations with the reference. Voita et al. (2018) used attention scores which show a tendency of Transformer-based context-aware models to do anaphora resolution. However, Müller et al. (2018) report moderate improvements of the model on their pronoun test set. In order to provide a comprehensive evaluation of our approach, we use BLEU, the pronoun challenge set from Müller et al. (2018), and F_1 score for the ambiguous English pronoun "it" based on alignment.

Previous work on curriculum learning for MT (Kocmi and Bojar, 2017; Zhang et al., 2018b; Wang et al., 2018) proposed methods which feed easier samples to the model first and later show more complex sentences. However, their focus is on improving convergence time while providing limited success on improving translation quality. In contrast with their work, we train models to better handle discourse-level phenomena.

3 Model

We use the Transformer (Vaswani et al., 2017) as a baseline and implement a context-aware model on top of it using Sockeye[1] (Hieber et al., 2018). The main and context sentence encoders are shared up until the penultimate layer, while the last encoder layers are separate. Since the initial layers are shared, the context sentence is marked with a special token so that the encoder knows when a context sentence is being encoded.

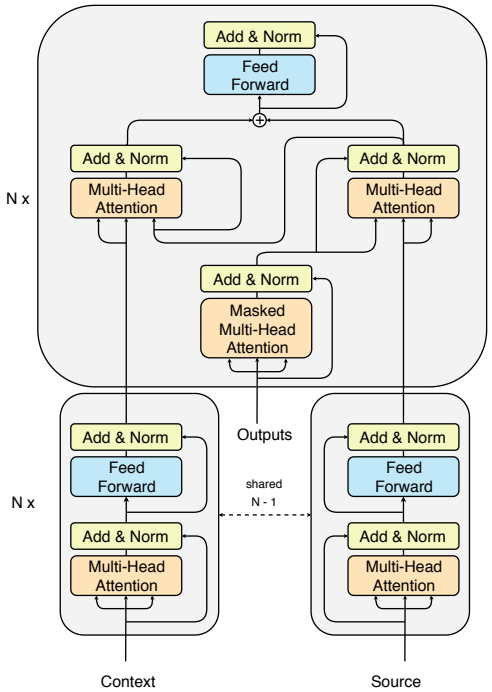

Figure 1: Context-aware model

The decoder layer is based on the standard Transformer decoder. It contains sublayers for

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

self-attention over the target and multi-head attention (MHA) over the encoded main sentence representation. We further introduce a MHA sublayer over the context representation. The output of the main sentence MHA is used as a query for the MHA over the context which represents the keys and the values. The MHA maps the queries and the keys in order to produce attention weights to score the values. In this way, the context MHA is conditioned on what has been generated until the given time step and on the main sentence. This helps the model to decide where to pay attention to in the context. The outputs of the MHA over the main and context sentences are merged using a gated sum which enables the model to control the flow of information between the main and context sentence. Finally, we apply a feed-forward network. All embeddings in the model including the context embeddings are shared. For further details on the Transformer, we refer to (Vaswani et al., 2017).

4 Curriculum Learning Method

The proposed approach leverages discourse-specific oracles (Stojanovski and Fraser, 2018) in a curriculum learning setting to improve the performance of context-aware models in terms of anaphora resolution on English→German translation. Antecedents to anaphoric pronouns are often in previous sentences. We therefore bias the model to pay more attention to the context when translating pronouns, thus enabling it to do better anaphora resolution. This is facilitated by providing oracle information in the context. Subsequently, oracles are gradually removed with the final result that we finish with a model which is not dependent on oracle information, but which knows that anaphoric pronouns are likely to be resolved by looking at previous sentence context.

4.1 Obtaining oracles

We modify the dataset with oracle information by extracting all pronouns from a reference target sentence and adding them to the corresponding source context sentence. In this work, we only use the previous source sentence. To some extent this is sufficient as in many cases antecedents are relatively close to the corresponding anaphoric pronouns. Distance-based statistics of antecedents in the challenge set (Müller et al., 2018) support this. Previous work (Miculicich et al., 2018; Zhang et al., 2018a) has shown that larger context does not provide for significant improvements, but these works have not conducted a tailored evaluation of anaphora resolution with regards to machine translation. We leave consideration of further context sentences for future work.

context sentence
The woman told a joke$^{[\text{masculine}]}$.
source sentence
It was really funny.
oracle sentence
The woman told a joke. er$^{[\text{masculine}]}$ [SEP] <PRON> It was really funny.
target sentence
Er war wirklich lustig.

Table 1: Oracle example. [SEP] - context separator; <PRON> - pronoun mark token. Glosses for presentation purposes only.

The method of obtaining oracles works as follows. For a given source sentence and reference target sentence we mark all source side pronouns, and extract all target side pronouns and insert them in the context sentence. We mark the pronouns by adding a special token <PRON> before the pronoun. Note that we always mark source side pronouns in the main sentence only (the sentence being translated). In a pure oracle setting, there is no need to mark all source side pronouns. In some sentence pairs, there are no pronouns on the target side and therefore there is no need to mark source pronouns since they don't need to be explicitly translated. However, our goal is through curriculum learning to end up with a non-oracle model and any oracle knowledge is undesirable. The extracted target side pronouns (taken from the main target sentence) are simply inserted at the end of the context sentence.

Consider the example in Table 1. [SEP] is a token marking the end of the context and beginning of the main sentence. The glosses in the examples are not in the actual data samples and are just used for presentation purposes in the paper. In the example in Table 1 we can see that the source sentence contains a pronoun "it" and the target sentence contains a pronoun "er". From the example, it is obvious that "er" is a translation of "it" and "it" is a anaphoric pronoun whose antecedent is present in the previous sentence, namely, "joke".

Given the main sentence alone, it is impossible to determine the appropriate gender of the third person singular pronoun in German. A baseline model will fall back to the data driven prior which tends to be the neuter form "es". However, the translations of "joke" in German, which commonly are "Witz" or "Scherz" are both masculine.

By inserting the correct information to resolve the gender in the context, we bias the model to pay more attention to the context when translating pronouns. This will not be of importance for some English pronouns which are gender independent (e.g., "I"), but it should be helpful for gender-ambiguous pronoun translations such as the English "it" (which must be translated consistently with the antecedent).

4.2 Training curriculum

The training curriculum is designed in order to make use of the oracle information. Previous work has focused on gradually increasing the complexity of the data being fed into a given model. Our approach is conceptually similar in the sense that initially the information for proper anaphora resolution is made explicit. Oracle reference pronouns in the context enable this. It does not necessarily mean that the data examples are less complex, but the model does not need to learn complex pronoun-antecedent relationships at the beginning.

An overview of the general curriculum training steps are:

- train a non-context-aware baseline Transformer model
- use the parameters of the baseline model to initialize the non-context parameters of the context-aware Transformer model
- train the context-aware model with an oracle dataset (gold-standard pronouns in the context)
- fine-tune the model with a dataset where the percentage of oracle samples is gradually lowered
- fine-tune the last model with a non-oracle dataset

We first train a baseline model without giving access to contextual information. The trained parameters are used to initialize the context-aware models (sublayers of the network dealing with context are randomly initialized). The following step is obtaining oracles for each sample in the dataset and training a model on that data. Resolving the gender of anaphoric pronouns in such a setting is easy. When the model encounters the special token marking a source side pronoun it will learn to look at the context since the gold standard information is there. We specifically put the oracle reference pronouns in the context in order to bias the model to pay attention to the context.

However, applying this model straightforwardly in a realistic setting is not possible because it is biased to rely on the gold standard pronouns. As a result, the next step is fine-tuning this model with context which does not contain the gold standard pronouns, but still has marked source side pronouns. In this way, we still bias the model to look at the context when translating pronouns. However, it is possible it will be difficult for the model to handle the significant change between fine-tuning steps.

As a result, we studied extending the training curriculum with intermediate steps. The initial oracle model is fine-tuned with a dataset where 75% of the samples have oracles. For the remaining samples, we keep the previous sentence and remove the oracle signals. In consecutive steps, we propose to fine-tune the model with a 50% and 25% oracle dataset. We hoped that this would ease the transition and encourage the model to combine the oracle information with the previous sentence. In the final step, we train a model with the previous sentence as context. This step is necessary as the model is still biased to look for the gold standard pronouns. However, we experimentally show that better results are obtained with fewer steps using a low percentage of oracles.

5 Experimental Setup

Following Müller et al. (2018), we conduct experiments on English→German WMT17 data and use newstest2017 and newstest2018 as test sets in addition to the pronoun challenge set. In terms of preprocessing, we tokenize and truecase the data and apply BPE splitting (Sennrich et al., 2016) with 32000 merge operations. We remove all samples where the source, target or context sentence has length over 50. We train small Transformer models as outlined in Vaswani et al. (2017) with 6 encoder and decoder layers. The source code for

our models is publicly available [2].

We report mean scores across ten consecutive checkpoints with the lowest average perplexity on the development set (Chen et al., 2018). BLEU scores are computed on detokenized text. Evaluation of pronoun translation is done using two separate metrics. First, we use the challenge set provided by Müller et al. (2018) and report the overall pronoun accuracy. We refer to this metric as challenge set accuracy. The other metric is an F_1 score for "it", which we refer to as reference F_1. We predict translations and then compute micro-average F_1 for "it", using an alignment of the test set input to the reference. We compute alignments using *fastalign* (Dyer et al., 2013). We use all of the training, development and test data for the computation of the alignments. The evaluation was done using the script from Liu et al. (2018).

6 Results

6.1 Baseline

We train a strong Transformer-based baseline which obtains different results than the baseline in Müller et al. (2018). We achieve higher BLEU scores and also observe different challenge set accuracy for the different pronouns, even though the overall score of 47% is similar. All context-aware models are initialized from this strong baseline. We create two setups, i) an initial setup where we train context-aware models with a high learning rate and ii) an improved setup where we train models with a low learning rate.

6.2 Initial setup

As a context-aware baseline (ctx-base), we train a model using the previous source sentence without access to gold standard pronouns. We assumed that a low learning rate could prevent the context-aware models to significantly change the baseline prior pronoun distribution. As a result, we use a high learning rate (10^{-4}) in the fine-tuning step. Training the context-aware baseline for 200K updates provides a small increase in BLEU on newstest, as shown in Table 2. However, large improvements are obtained on the subtitles challenge set. We attribute this to the higher dependency on the context in subtitles which benefits from the increased capability of the context-aware model to diverge from the baseline.

[2] https://www.cis.uni-muenchen.de/~dario/projects/curriculum-oracles

	nt17	nt18	challenge
baseline	26.9	40.0	21.7
ctx-base*	27.0†	40.2‡	**22.6**†
ctx-base**	27.2†	**40.4**†	22.0†
pron-25→pron-0*	26.9	39.9	**22.6**†
pron-25→pron-0**	**27.4**†	40.2	22.2†

Table 2: BLEU scores. * - initial learning rate is 10^{-4}, ** - lr=10^{-5}. ctx-base: context-aware baseline, pron-{0,25,50,75}: percentage of samples with oracles. Each pron-{0,25} model fine-tuned for 140K updates. †- improvements statistically significant based on paired bootstrap resampling with p-value < 0.01; ‡- p-value < 0.05

	nt17	challenge
baseline	65.8	36.0
ctx-base*	**67.1**	**45.3**
ctx-base**	65.1	38.1
pron-25→pron-0*	65.2	45.1
pron-25→pron-0**	65.5	40.2

Table 3: Reference F_1 for "it" on newstest2017 and the pronoun challenge set. Notation as in Table 2

However, our curriculum learning approach does not affect performance in this setting. Figure 2 shows that the context-aware baseline achieves 57% challenge set accuracy and the curriculum learning approach only manages to match the score. Figure 2 further depicts that using a high number of oracle pronouns in the dataset decreases performance and that fine-tuning these models with a lower percentage of oracles is not useful. For example, fine-tuning a 25% oracle (pron-25) from the baseline is better than fine-tuning from a 50% oracle considering equal training time. The other oracle settings perform similarly. As a result, the full training curriculum from 100% gradually to 0% oracles is not justified both in terms of computation time or performance. Fine-tuning pron-25→pron-0 for a longer amount of time improved to 58%, but we omit it from the figure since we did not train ctx-base for a comparable amount of time. In terms of reference F_1, shown in Table 3, the context-aware baseline achieves large improvements in comparison to the baseline, both on newstest2017 and the challenge set, but our proposed method fails to increase performance.

6.3 Improved setup

Training context-aware models with a high learning rate improves overall translation quality on subtitles, but not on newstest. The high learning rate allows the model to diverge from the well-

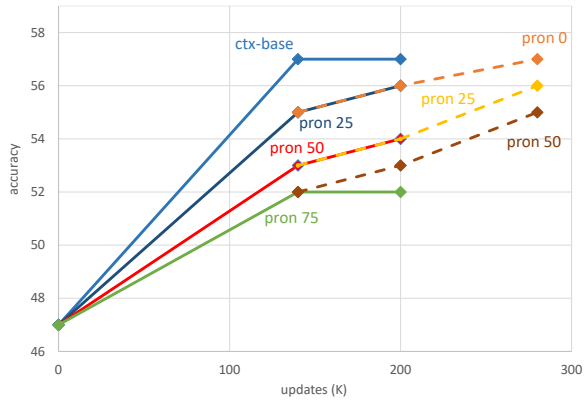

Figure 2: Challenge set accuracy. Full lines show fine-tuning from the baseline and dashed lines from a previous oracle model. Fine-tuning with a lr=10^{-4}.

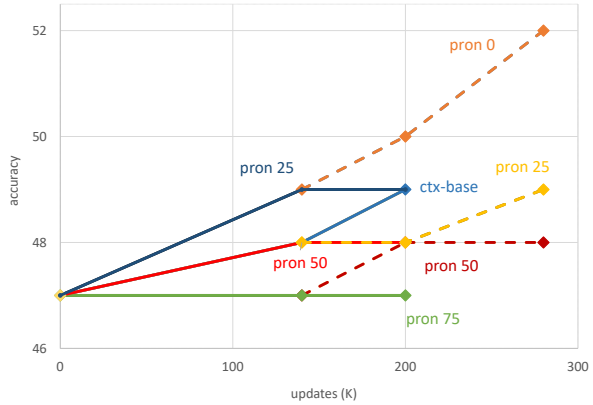

Figure 3: Challenge set accuracy. lr=10^{-5}.

optimized baseline and this affects performance. We therefore decided to train models with a low learning rate of 10^{-5}. In this setup, the ctx-base improves on newstest and subtitles by 0.3 or 0.4 BLEU. The gains in BLEU are smaller than the ones reported by Müller et al. (2018), but we compare against a stronger baseline.

Unfortunately, performance on pronoun translation is lower. Figure 3 shows that ctx-base improves challenge set accuracy only to 49%. However, in this experimental setup, our curriculum learning approach proved to be effective if we start-off the training curriculum with a lower percentage of oracles. If we train a context-aware baseline (ctx-base) for 200K updates, we get lower performance (49%) than training a 25% oracle (pron-25) for 140K updates and then fine-tuning with a 0% oracle (pron-25→pron-0) for 60K updates (50%). Fine-tuning this model for 140K updates further improves to 52%. Table 3 shows that it is also helpful on reference F_1, providing a 2.1 improvement over the 38.1 F_1 the ctx-base achieved on the challenge set.

All experiments show that fine-tuning with a high learning rate helps with pronoun translation, but does not benefit from the curriculum learning and lags behind training with a low learning rate in terms of BLEU. Therefore, we conclude that the curriculum learning is useful when improvements on anaphora resolution are desirable at no detrimental cost to overall translation quality.

6.4 Anaphora resolution analysis

We use the challenge set (Müller et al., 2018) to do a more detailed analysis of the models. We previously gave a high-level overview of the models' performance on the challenge set by only reporting the total score. The total score represents the overall accuracy, meaning the percentage of correctly scored examples. However, the challenge set is more comprehensive and offers a more detailed look at different aspects of anaphora resolution. As with the previous results, we report mean scores across ten consecutive checkpoints. We also report the standard deviation since we observed some degree of variance in the results depending on the experimental setup. Each fine-tuning step from the curriculum learning is ran for 140K updates.

6.4.1 Reference pronoun accuracy

Table 4 shows the overall and per-pronoun accuracy. Comparing our Transformer baseline to the one from Müller et al. (2018) showed that our baseline is stronger in terms of translation quality as measured by BLEU. However, in terms of pronoun accuracy as measured by the challenge set, the performance is the same with differences on the per-pronoun accuracy.

Table 4 also shows the detail scores for the context-aware baselines and the curriculum setup where we first train with a 25% oracle and fine-tune with a 0% oracle. Scores are provided for both fine-tuning with a low and high learning rate. The high learning rate context-aware baseline obtains 0.37 on "er", 0.44 on "sie" and a high 0.92 on "es". The curriculum experiment pron-25→pron-0 has similar scores with a lower accuracy on "sie".

The detailed scores also show how the low learning rate models perform. Both, the context-aware baseline and pron-25→pron-0 improve over the baseline. Another aspect that speaks for using fine-tuning with low learning is stability of results. Although the high learning rate models improve

	total	er	sie	es
baseline	0.47 ± 0.003	0.20 ± 0.005	0.32 ± 0.011	0.89 ± 0.005
ctx-base*	0.57 ± 0.007	0.37 ± 0.014	0.44 ± 0.019	0.92 ± 0.005
ctx-base**	0.49 ± 0.003	0.23 ± 0.006	0.35 ± 0.010	0.90 ± 0.004
pron-25→pron-0*	0.57 ± 0.013	0.37 ± 0.027	0.42 ± 0.032	0.92 ± 0.009
pron-25→pron-0**	0.52 ± 0.005	0.26 ± 0.010	0.38 ± 0.010	0.91 ± 0.001

Table 4: Challenge set accuracy for each pronoun. Notation as in Table 2

	intrasegmental	external
baseline	0.73 ± 0.005	0.41 ± 0.004
ctx-base*	0.74 ± 0.011	0.53 ± 0.009
ctx-base**	0.73 ± 0.006	0.43 ± 0.004
pron-25→pron-0*	0.74 ± 0.016	0.53 ± 0.014
pron-25→pron-0**	0.74 ± 0.004	0.46 ± 0.005

Table 5: Challenge set accuracy based on location of antecedent. Notation as in Table 2

fast on anaphora resolution, they are relatively unstable and exhibit fair amount of variance on the challenge set evaluation. This was to some extent observed on BLEU scores as well, but it is less pronounced. A difference in results across different checkpoints is especially observed on "er" and "sie". The experiments with a low learning rate exhibit variance on par with the baseline. This shows that reporting results on the challenge set needs to be carefully executed.

6.4.2 Antecedent location

The challenge set also provides a way of evaluation based on the location of the antecedent. There are two categories, intrasegmental and intersegmental or external. The intrasegmental means that the antecedent is within the main sentence. External refers to examples where the antecedent is in a previous sentence. It is unsurprising to observe that all models, including non-context and context-aware models perform similarly on the intrasegmental score and most of the improvements come from looking at the context, which is what the external score in Table 5 shows.

6.4.3 Antecedent distance

Table 6 shows scores based on the distance of the antecedent. The distance can be 0 (in the main sentence), 1 (in the first previous sentence) or larger. In this work, we only use the first previous sentence, so the results for a distance of 2, 3 or larger are for comparison with previous work. It is again unsurprising that performance does not substantially differ for 2, 3 or >3 since our models do not have direct access to those sentences. Any difference in results most likely comes from changing the data driven prior of the baseline. All improvements of the context-aware models come from examples where the antecedent is in the first previous sentence. We see that pron-25→pron-0 with a low learning rate obtains high improvements of 0.07 in comparison to the baseline.

6.5 Attention analysis

The model proposed in this work incorporates the contextual representation in each layer in the decoder. This raises the question what layers are responsible for finding the appropriate information for anaphora resolution. Unlike previous RNN-based encoder-decoder architectures which have a single attention mechanism, the Transformer is implemented using multi-head attention. As a result, we first average the attention scores across all attention heads and then visualize the scores.

We do a detailed analysis for separate decoder layers. Figure 4, Figure 5, Figure 6 and Figure 7 show the attention scores from the first, second, third and last layer. The attention scores are from pron-25→pron-0 with a low learning rate.

All context sentences are preceded by the <ctx> token. An interesting phenomena which was also observed in Voita et al. (2018) is that this special token is paid a substantial amount of attention. They interpret this as a way for the model to ignore the context when not needed.

The visualizations show that this is not the case for our model. We observe that the model takes advantage of the fact that the context is used in multiple layers. In the first 3 layers, the models generally pay the highest attention to the appropri-

	0	1	2	3	>3
baseline	0.73 ± 0.005	0.37 ± 0.005	0.47 ± 0.003	0.50 ± 0.004	0.69 ± 0.010
ctx-base*	0.74 ± 0.011	0.54 ± 0.011	0.47 ± 0.005	0.51 ± 0.008	0.72 ± 0.009
ctx-base**	0.73 ± 0.006	0.40 ± 0.005	0.47 ± 0.002	0.50 ± 0.004	0.69 ± 0.008
pron-25→pron-0*	0.74 ± 0.016	0.53 ± 0.017	0.46 ± 0.005	0.50 ± 0.010	0.71 ± 0.008
pron-25→pron-0**	0.74 ± 0.004	0.44 ± 0.007	0.46 ± 0.003	0.50 ± 0.004	0.69 ± 0.004

Table 6: Challenge set accuracy based on distance of antecedent. Notation as in Table 2

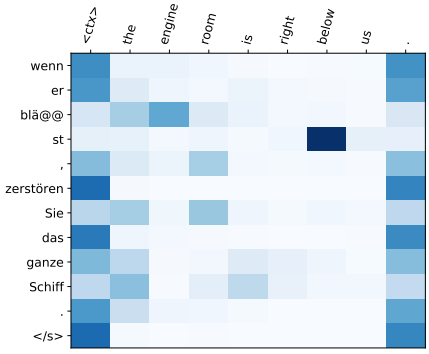

Figure 4: Context attention layer 1

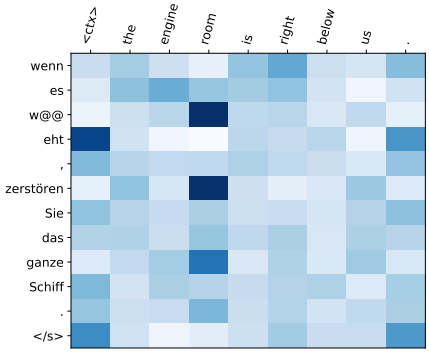

Figure 5: Context attention layer 2

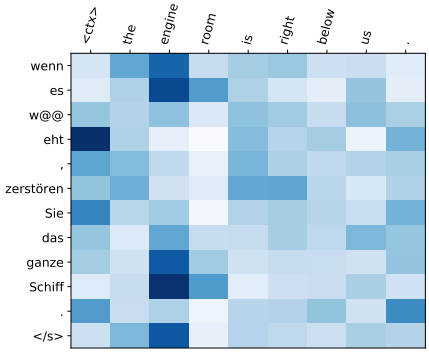

Figure 6: Context attention layer 3

ate noun, but a lot of attention is paid to irrelevant parts of the previous sentence. However, we see that the attention sharpens in the last layer and the attention over the context mostly focuses on the appropriate tokens. The example we show here is

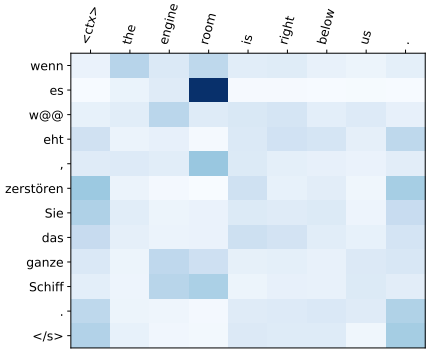

Figure 7: Context attention layer 6

a negative example as the correct German pronoun is “er” while the model generated “es”[3].

In contrast, we didn’t observe the same behavior from pron-25→pron-0 with a high learning rate. This model indeed seemed to consistently put attention on the context special token and at the end of the sentence. Attention was paid to the antecedent in the decoder layers by target pronouns, but also by other words in some cases, leading us to assume that the gender information was passed through the decoder. We also assumed that the context special token to some extent represents a summarized representation of the context sentence and contains some gender information. Masking this token when feeding the context encoder representation to the decoder leads to lower results on the challenge set. We leave a more detailed examination of this assumption for future work.

6.5.1 Commonly attended words

We further investigate what words are most commonly attended to by the reference pronouns “er”, “sie”, “es”. We simply compute the total attention score paid to a given context source token by one of the pronouns. We then normalize the scores based on the frequency of the given word.

[3]The translation of engine room in German is a compound word (Maschinenraum or Motorraum) and the gender is inferred from the second part, namely, “Raum”. “Raum” is masculine in German, but a more common translation of “room” is “Zimmer” whose gender is neuter.

er	SU@@, Cube, Var@@, Max, ulf, tunnel, text, mur@@, schedule, passport, Jean, painting, bug, President, enemy, Ring, 400@@, temple, spell, state, Frank@@, Key, Cra@@, container, Doctor, Tony, recognized
sie	covers, Body, marble, painting, Machine, church, obviously, Lin@@, gar@@, decision, chamber, party, grie@@, Ara@@, hat@@, humanity, Enterprise, identity, Box, eventually, force, teeth, technology, Anne, tro@@, milk, policy
es	palace, fantastic, Ver@@, Jack@@, Board, article, museum, meeting, seed, So@@, gold, sample, technique, beef, satellite, Dal@@, virus, promise, piano, Jesus, Mac@@, motion, adventure, sounds, Cav@@, match, Ford

Table 7: Frequency based attention analysis

Since we are working on the BPE level, it is sometimes difficult to determine whether the attention score is meaningful, but it gives some indication whether the models are working correctly.

We show the most attended words from the pron-25→pron-0 with a low learning rate. Context words which appeared in a sentence containing a pronoun less than 5 times were removed in order to reduce the probability that some words are attended by chance. We only use the lowercase versions of the pronouns since "Sie" in German can also refer to the polite version of "you" and it cannot easily be disambiguated. We show the source tokens in Table 7. A detailed automatic analysis is problematic because English words can have multiple translations in German and sometimes those translations have different genders. We manually looked at common German translations of the tokens in Table 7. We noticed that in many cases the gender of the translation corresponds to the gender of the pronoun. We also looked at the non-BPE-split tokens and mapped them to German words using the MUSE English-German bilingual dictionary (Lample et al., 2018). We then looked at the gender of the German translations and how often it corresponds to the pronoun gender. The pron-25→pron-0 model performed better compared to the context-aware baseline, meaning a higher percentage of the German translations had gender corresponding to the gender of the pronoun. We leave a more detailed manual evaluation for future work.

7 Conclusion

We devised a curriculum learning approach making use of oracle information to improve anaphora resolution in NMT. Tailoring the data and training curriculum to anaphora resolution is beneficial and can achieve gains against a context-aware baseline. We observed that fine-tuning with low learning rates when applying our curriculum learning method provides a good compromise between overall translation quality and pronoun accuracy. Our method works best with a small number of fine-tuning steps employing smaller percentages of oracles. Our work is a focused contribution showing that curriculum training can be used to improve translation accuracy beyond a starting baseline given oracle information. Our experiments show that using a small learning rate during training is important to obtain improvements.

One aspect of our work that we do not explore is different ways of generating the oracle datasets. We always randomly sampled the sentences that are to be modified with the reference target side pronouns. Future work can investigate more informed ways of creating the oracle datasets. The benefit of this direction is that creating several different random samples of the oracle datasets could provide for more diverse models. This can be very useful for ensembling where larger variety between models is desirable. One could imagine that the variety in the models introduced by this approach is going to be more useful than if we simply train different baselines, context-aware or not.

It is also promising to try our method with other discourse-level phenomena that have easily obtainable oracles. Coherence and cohesion are important aspects of machine translation and improving on those discourse-level phenomena is still challenging for sentence-level models.

Acknowledgments

We would like to thank Dan Bikel and the anonymous reviewers for their valuable input. 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

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Chen, Mia Xu, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86. Association for Computational Linguistics.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA, March. Association for Machine Translation in the Americas.

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

Kocmi, Tom and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386. INCOMA Ltd.

Kuang, Shaohui and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 607–617, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kuang, Shaohui, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Liu, Frederick, Han Lu, and Graham Neubig. 2018. Handling homographs 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 1336–1345.

Maruf, Sameen and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284. Association for Computational Linguistics.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954. Association for Computational Linguistics.

Müller, Mathias, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 61–72, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Stojanovski, Dario and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 49–60, Brussels, Belgium, October. Association for Computational Linguistics.

Tiedemann, Jörg and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Tu, Zhaopeng, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Voita, Elena, 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.

Wang, Longyue, 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.

Wang, Rui, Masao Utiyama, and Eiichiro Sumita. 2018. Dynamic sentence sampling for efficient training of neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 298–304. Association for Computational Linguistics.

Zhang, Jiacheng, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018a. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542. Association for Computational Linguistics.

Zhang, Xuan, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018b. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.