Modeling Contextual Information in Neural Machine Translation

Dario Stojanovski

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Dario Stojanovski

München, den 29. April 2021

Erstgutachter: Prof. Dr. Alexander Fraser
Zweitgutachter: Prof. Dr. Philipp Koehn
Drittgutachter: Prof. Dr. Rico Sennrich

Tag der Einreichung: 29.04.2021
Tag der mündlichen Prüfung: 30.06.2021

Eidesstattliche Versicherung
(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig ohne unerlaubte Beihilfe angefertigt ist.

München, den 12.07.2021

Dario Stojanovski

Acknowledgments

First and foremost, I would like to thank my advisor Alexander Fraser for providing invaluable support throughout my PhD journey. He made a significant effort in making me become a better researcher and guided me how to do rigorous science. We always had fruitful discussions about my ongoing research. He also provided personal support in some of the more challenging times of my PhD work.

During my time at CIS, I worked and collaborated with many wonderful people. I would like to pay special thanks to my office mates, Matthias Huck and Alexandra Chronopoulou. I am also glad we had the chance to collaborate on research projects. I also like to thank the other members of the Machine Translation team at CIS, Helmut Schmid, Fabienne Braune, Viktor Hangya, Jindřich Libovický and Denis Peskov whom I also had the fortune to collaborate with and learn from.

CIS is a large group. First, I am happy I had the chance to work in the same group with Hinrich Schütze and to learn from him at our numerous reading groups and PhD seminars. I am glad I had the chance to briefly overlap with Yadollah Yaghoobzadeh, Heike Adel, Wenpeng Yin and Katharina Kann who were very kind, made me feel welcome in the group and provided helpful advice in the writing of the thesis. I would also like to thank Benjamin Roth, Philipp Dufter, Ehsaneddin Asgari, Masoud Jalili Sabet, Nora Kassner, Nina Poerner, Martin Schmitt, Dietrich Trautmann, Mengije Zhao, Marina Sedinkina and many others whose research or non-research chats made my time at CIS especially memorable. Special thanks to Peggy Hobmaier and Thomas Schäfer for all the help with the administrative and IT support issues I encountered, but for the fun chats as well.

I was fortunate to be able to do an internship at Amazon Research where I learned about the practical applications of our scientific research. I am very grateful for the close collaboration I had with Tobias Domhan and Felix Hieber.

Finally, none of this would have been possible without the never-ending support of my wife Nikolina, my parents Igor and Zorica and my brother Nikola. I

am immensely grateful for the all the love and support from my wife and especially thankful for her putting up with me in the days before conference deadlines. While geographically apart, my parents and brother were also always there for me.

Abstract

Machine translation has provided impressive translation quality for many language pairs. The improvements over the past few years are largely due to the introduction of neural networks to the field, resulting in the modern sequence-to-sequence neural machine translation models. NMT is at the core of many large-scale industrial tools for automatic translation such as Google Translate, Microsoft Translator, Amazon Translate and many others.

Current NMT models work on the sentence-level, meaning they are used to translate individual sentences. However, for most practical use-cases, a user is interested in translating a document. In these cases, an MT tool splits a document into individual sentences and translates them independently. As a result, any dependencies between the sentences are ignored. This is likely to result in an incoherent document translation, mainly because of inconsistent translation of ambiguous source words or wrong translation of anaphoric pronouns. For example, it is undesirable to translate "bank" as a "financial bank" in one sentence and then later as a "river bank". Furthermore, the translation of, e.g., the English third person pronoun "it" into German depends on the grammatical gender of the English antecedent's German translation.

NMT has shown that it has impressive modeling capabilities, but is nevertheless unable to model discourse-level phenomena as it needs access to contextual information. In this work, we study discourse-level phenomena in context-aware NMT. To facilitate the particular studies of interest, we propose several models capable of incorporating contextual information into standard sentence-level NMT models. We direct our focus on several discourse phenomena, namely, coreference (anaphora) resolution, coherence and cohesion. We discuss these phenomena in terms of how well can they be modeled by context-aware NMT, how can we improve upon current state-of-the-art as well as the optimal granularity at which these phenomena should be modeled. We further investigate domain as a factor in context-aware NMT. Finally, we investigate existing challenge sets for anaphora

resolution evaluation and provide a robust alternative.

We make the following contributions:

i We study the importance of coreference (anaphora) resolution and coherence for context-aware NMT by making use of oracle information specific to these phenomena.

ii We propose a method for improving performance on anaphora resolution based on curriculum learning which is inspired by the way humans organize learning.

iii We investigate the use of contextual information for better handling of domain information, in particular in the case of modeling multiple domains at once and when applied to zero-resource domains.

iv We present several context-aware models to enable us to examine the specific phenomena of interest we already mentioned.

v We study the optimal way of modeling local and global context and present a model theoretically capable of using very large document context.

vi We study the robustness of challenge sets for evaluation of anaphora resolution in MT by means of adversarial attacks and provide a template test set that robustly evaluates specific steps of an idealized coreference resolution pipeline for MT.

Zusammenfassung

Die maschinelle Übersetzung erreicht für viele Sprachpaare eine beeindruckende Übersetzungsqualität. Die Fortschritte der letzten Jahre sind größtenteils auf die Einführung neuronaler Netze, insbesondere moderner Sequenz-zu-Sequenz-Modellen, in diesem Bereich zurückzuführen. Neuronale maschinelle Übersetzung (NMÜ) ist das Kernstück viel genutzter kommerzieller Applikationen wie Google Translate, Microsoft Translator oder Amazon Translate.

Aktuelle NMÜ-Modelle arbeiten auf Satzebene. Das heißt, sie werden für die Übersetzung einzelner Sätze verwendet. In den meisten praktischen Anwendungsfällen will ein Nutzer jedoch ein Dokument zu übersetzen. In diesem Fall teilt ein NMÜ-Modell ein Dokument in einzelne Sätze und übersetzt diese unabhängig voneinander. Folglich werden etwaige Abhängigkeiten zwischen den Sätzen ignoriert. Dies kann zu einer inkohärenten Übersetzung des Dokuments führen, vor allem aufgrund inkonsistenter Übersetzung mehrdeutiger Wörter oder falscher Übersetzung anaphorischer Pronomen. Zum Beispiel ist es inkonsistent, "Bank" in einem Satz als Bank im Sinne von Finanzinstitut und im nächsten Satz als "Sitzbank" zu übersetzen. Ebenso hängt die Übersetzung des englischen Pronomens "it" im Deutschen vom Geschlecht der deutschen Übersetzung des englischen Antezedens ab.

NMÜ hat gezeigt, dass sie beeindruckende Modellierungsfähigkeiten hat. Dennoch ist sie nicht in der Lage, Phänomene auf Diskursebene zu modellieren, da der Zugang zu Kontextinformationen fehlt. In dieser Arbeit untersuchen wir Phänomene auf Diskursebene in kontextsensitiver NMÜ. Wir entwickeln mehrere Modelle, die in der Lage sind, kontextuelle Informationen in Standard NMÜ-Modelle auf Satzebene zu integrieren. Dabei richten wir unseren Fokus auf mehrere Diskursphänomene, nämlich die Auflösung von Koreferenzen (Anaphern), Kohärenz und Kohäsion. Wir diskutieren diese Phänomene im Hinblick darauf, wie gut sie durch kontextabhängige NMÜ modelliert werden können, wie wir den aktuellen Stand der Technik verbessern können und in welcher Granularität diese

Phänomene modelliert werden sollen. Weiterhin untersuchen wir die Domäne in kontextbezogener NMÜ. Zuletzt untersuchen wir Datensätze für die Bewertung der Anapherauflösung und entwickeln eine robuste Alternative.

Der Beitrag dieser Arbeit ist:

i Wir untersuchen die Bedeutung von Koreferenzenauflösungen (Anaphern) und die Kohärenz in kontextsensitiver NMÜ, indem wir Orakelinformationen für diese Phänomene nutzen.

ii Wir entwickeln eine Methode zur Verbesserung der Koreferenzenauflösungen, die auf dem Lernen von Curricula basiert, welches wiederum von menschlichem Lernen inspiriert ist.

iii Wir untersuchen die Verwendung von Kontextinformationen für eine bessere Verarbeitung von Domäneninformationen, insbesondere im Fall der gleichzeitigen Verarbeitung mehrerer Domänen und bei der Verarbeitung von Domänen ohne jegliche Trainingsdaten.

iv Wir stellen mehrere kontextabhängige Modelle vor, die es uns ermöglichen, die erwähnten Phänomene zu untersuchen.

v Wir untersuchen die Modellierung von lokalem und globalem Kontext und stellen ein Modell vor, das theoretisch in der Lage ist, besonders großen Dokumentenkontext zu nutzen.

vi Wir untersuchen die Robustheit der Datensätze für die Evaluierung der Anapherauflösung in maschineller Übersetzung mithilfe von “adversarial” Attacken und stellen ein Template-Test-Set zur Verfügung, das Schritte einer idealisierten Pipeline für die Koreferenzauflösung in NMÜ robust evaluiert.

Contents

Publications and Declaration of Co-Authorship

Chapter 2

Chapter 2 corresponds to the following publication:

> Dario Stojanovski, Alexander Fraser; **Coreference and Coherence in Neural Machine Translation: A Study Using Oracle Experiments**; Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers (Brussels, Belgium, October, 2018), pages 49–60.

I regularly discussed this work with my advisor, but I conceived of the original research contributions and performed implementation and evaluation. I wrote the initial draft of the article and did most of the subsequent corrections. My advisor assisted me in improving the draft.

Chapter 3

Chapter 3 corresponds to the following publication:

> Dario Stojanovski, Alexander Fraser; **Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning**; Proceedings of Machine Translation Summit XVII Volume 1: Research Track, (Dublin, Ireland, August, 2019), pages 140–150.

I regularly discussed this work with my advisor, but I conceived of the original research contributions and performed implementation and evaluation. I wrote the initial draft of the article and did most of the subsequent corrections. My advisor assisted me in improving the draft.

Chapter 4

Chapter 4 corresponds to the following publication:

> Dario Stojanovski, Alexander Fraser; **Addressing Zero-Resource Domains Using Document-Level Context in Neural Machine Translation**; Proceedings of the Second Workshop on Domain Adaptation for NLP, (Kyiv, Ukraine, April, 2021), pages 80–93.

I regularly discussed this work with my advisor, but I conceived of the original research contributions and performed implementation and evaluation. I wrote the initial draft of the article and did most of the subsequent corrections. My advisor assisted me in improving the draft.

Chapter 5

Chapter 5 corresponds to the following publication:

> Dario Stojanovski, Alexander Fraser; **Combining Local and Document-Level Context: The LMU Munich Neural Machine Translation System at WMT19**; Proceedings of the Fourth Conference on Machine Translation Volume 2: Shared Task Papers, (Florence, Italy, August, 2019), pages 400—406.

I regularly discussed this work with my advisor, but I conceived of the original research contributions and performed implementation and evaluation. I wrote the initial draft of the article and did most of the subsequent corrections. My advisor assisted me in improving the draft.

Chapter 6

Chapter 6 corresponds to the following publication:

> Dario Stojanovski, Benno Krojer, Denis Peskov, Alexander Fraser; **ContraCAT: Contrastive Coreference Analytical Templates for Machine Translation**; Proceedings of the 28th International Conference Computational Linguistics, (Barcelona, Spain, December, 2020), pages 4732—4749.

This work is the result of a collaboration. Benno Krojer, Denis Peskov and I contributed in equal parts. I conceived of the original research idea. I trained

the machine translation models and worked on the "augmentation" part of the work. Benno Krojer and I collaborated on the "adversarial attack" part of the work. I regularly discussed the "templates" part of the work with Benno Krojer and Denis Peskov. I wrote the initial draft of the introduction, related work, model details and augmentation. Benno Krojer wrote the initial draft of the adversarial attacks and templates. Benno Krojer, Denis Peskov and I contributed equally in the subsequent corrections and improvements of these sections. We regularly discussed this work with my advisor. My advisor assisted us in improving the draft.

In parallel to the publications related to the thesis, I have authored or co-authored the following publications:

> Matthias Huck, Dario Stojanovski, Viktor Hangya, Alexander Fraser; **LMU Munich's Neural Machine Translation Systems at WMT 2018**; Proceedings of the Third Conference on Machine Translation Volume 2: Shared Task Papers, (Brussels, Belgium, November 2018), pages 648—654.
>
> Dario Stojanovski, Viktor Hangya, Matthias Huck, Alexander Fraser; **The LMU Munich Unsupervised Machine Translation Systems**; Proceedings of the Third Conference on Machine Translation Volume 2: Shared Task Papers, (Brussels, Belgium, November 2018), pages 513—521.
>
> Dario Stojanovski, Viktor Hangya, Matthias Huck, Alexander Fraser; **The LMU Munich Unsupervised Machine Translation System for WMT19**; Proceedings of the Fourth Conference on Machine Translation Volume 2: Shared Task Papers, (Florence, Italy, August 2019), pages 592—598.
>
> Alexandra Chronopoulou, Dario Stojanovski, Alexander Fraser; **Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT**; Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), (Online, November 2020), pages 2703—2711.
>
> Alexandra Chronopoulou, Dario Stojanovski, Viktor Hangya, Alexander Fraser; **The LMU Munich System for the WMT 2020 Unsupervised Machine Translation Shared Task**; Proceedings of the Fifth

Conference on Machine Translation Volume 2: Shared Task Papers, (Online, November 2020), pages 1084—1091.

Alexandra Chronopoulou, Dario Stojanovski, Alexander Fraser; **Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation**; Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), (Online, June 2021), pages 173—180.

München, 12.07.2021

Dario Stojanovski

Chapter 1

Introduction

Translation is of central importance in human interactions as it enables communication between speakers of different languages who do not speak a common language. Traditionally enabled by the presence of a bilingual speaker of two languages, the second half of the twentieth century has seen the rise of methods for automatic translation using software. The approach to automatize translation is referred to as machine translation (MT) and is widely used by professional translators and non-professionals alike due to many open platforms such as Google Translate, Amazon Translate and DeepL.

The majority of machine translation models today are based on neural networks. Neural machine translation (NMT) models use deep neural networks such as recurrent, convolutional or attention-based models and achieve remarkable performance on language pairs where large amounts of parallel data are available. Nevertheless, most NMT systems work on the sentence-level. They translate sentences individually and lack any ability to model the document-level dependencies that may arise. As a result, discourse-level phenomena are largely disregarded which provides for incoherent translations.

In this thesis, we study context-aware neural machine translation, a subfield of NMT focused on using contextual information. Context-aware NMT models attempt to remedy the challenges sentence-level models face in terms of document-level coherence. We present several context-aware MT models, better strategies to train them and how to better evaluate them. In the remainder of this section, we give a more detailed overview of machine translation, deep neural networks and their applicability to neural machine translation and finally discuss context-aware MT and the relevant discourse-level phenomena.

1.1 Machine Translation

Machine translation is the task of automatically translating human language from a source to a target language. It is a sub-field of natural language processing (NLP) which broadly deals with the automatic processing of natural language. Defined more narrowly, MT is a sequence-to-sequence task which takes a sequence of words as input and outputs a sequence of words, both of arbitrary length.

The output of an MT system should have some of the following desirable qualities. The generated target language should be adequate, meaning it should preserve the semantic and syntactic properties of the input language, and it should be fluent, meaning it should be effortlessly intelligible by a fluent speaker of the target language. The input is usually written text, but other modalities can be used as well, such as taking into account relevant visual aspects. Alternatively, an MT system can be used to translate speech into text or directly output target language speech. In this thesis, we focus on the text-to-text machine translation task. For the better part of the thesis, we use English as the source and German as the target language.

Machine translation can be tackled using different approaches. Some notable examples which have been used in the past include dictionary-based methods, rule-based machine translation (RBMT), statistical machine translation (SMT) and hybrid machine translation. Arguably, the least sophisticated approach is dictionary-based MT. The process is conducted by a simple lookup of the source words in a predefined dictionary of the source and target language. The method lacks any disambiguation capabilities and as expected, produces inarticulate target language. However, it has potential use-cases in very simple scenarios, such as translating single words in a given list or more interestingly, it may have some applications in the translation of very similar languages (Hajic, 1987). Rule-based MT (Johnson et al., 1985) is a method that works by using source and target language linguistic information to create rules for translation. The main advantage of these systems is that no parallel data is necessary. While this may seem appealing, these systems require extensive feature engineering and are in general very cumbersome to design.

Statistical machine translation (SMT) (Brown et al., 1993; Koehn, 2009) is a method that has been widely used prior to neural machine translation. One of the key characteristics of both approaches is the necessity of parallel data, sentences translated between two languages. Various SMT approaches have been proposed with the most popular one being phrase-based machine translation (PBMT)

(Koehn et al., 2003). In PBMT, the model translates phrases of a certain length. The model builds a phrase table where phrases are mapped one-to-one and uses parallel data to train a parameterized model using expectation maximization.

As of the writing of this thesis, the most widely used method for MT is using neural networks, which has become known as Neural Machine Translation (NMT). Unlike SMT where systems consist of multiple modules, NMT models are single unified architectures trained in an end-to-end fashion. NMT is based on neural networks, often feed-forward neural networks, recurrent neural networks and concepts such as attention, the details of which we discuss later in the thesis. All the modeling work and experiments in this thesis are conducted using NMT models.

One of the key requirements of NMT is the need for parallel sentences. NMT is data-driven. An NMT model has no prior knowledge, often referred to as inductive bias, on how to translate between any specific language pair. Nevertheless, an NMT model has inductive bias pertaining to abstract aspects of translation. NMT models process the source words independently, but also model the relationships between them. Different models do this differently and therefore exhibit different inductive biases. Furthermore, standard NMT models generate the translation one target word at a time, which has a resemblance to how humans generate translations. However, there is no explicit information encoded a priori of how to translate between a specific language pair. That information comes later as the model trains on parallel sentences. Parallel sentences are sentences in two languages that are a translation of each other. Subsequently, correlations between the source and target language are being learned and the model gains the ability to produce translations.

Through considerable effort, parallel data is available for a number of language pairs. Notably, the European parliament curates Europarl (Koehn, 2005), parallel translations of its proceedings in the languages of all European Union member states. Commercial interests can also lead to the creation of such datasets, for example, OpenSubtitles (Lison and Tiedemann, 2016), which contains a relatively high number of parallel sentences in many language pairs of movie subtitles. Despite these efforts, the sheer number of languages in the world leads to many language pairs having little or no parallel data. While this is a very important research topic, in this thesis, we focus on English→German translation for which large amounts of parallel data are available. Throughout the thesis, we use several different datasets, many of which are made available from WMT (Koehn and Monz, 2006) and OPUS (Tiedemann, 2012).

1.2 Deep Neural Networks

1.2.1 Fundamentals

One of the key areas of machine learning are neural networks which are statistical models for learning, very loosely based on the way biological neural networks work. Initial attempts in this area include the works of McCulloch and Pitts (1943); Hebb (1949); Rosenblatt (1958); LeCun et al. (1989) and Hinton (2007). Despite similarities to biological neural networks, they differ in many aspects such as artificial neural networks having a fixed topology or not being able to take into account the timing information of when neurons are firing.

A neural network consists of neurons or units with connections between them which transmit signals from one neuron to another. The connections between the neurons have weights which determine the strength of a signal passing the given connection. Neurons are arranged in layers and the layers are usually connected only to the preceding and the subsequent layer (apart from the obvious exceptions for the first and last layer in a network).

Feed-forward neural networks Feed-forward neural networks, sometimes also referred to as fully-connected neural networks, are a straightforward implementation of a neural network. The network consists of an input layer, an optional number of hidden layers and an output layer. The input layer's neurons are fully connected to the neurons of the first hidden layer, which are fully connected to the second hidden layer and so on until the last output layer.

A single layer in a neural network does the following computation:

$$f(x) = \sigma(Wx + b)$$

where x is a d-dimensional input. The neural network is parametrized by a weight matrix W and a bias vector b. For $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}^o$, $W \in \mathbb{R}^{d \times o}$ and $b \in \mathbb{R}^o$. Finally, σ is an activation function which determines the output of the layer. There are a number of possible activation functions that can be used. Some notable examples are the sigmoid function, Rectified Linear Units (ReLU) (Nair and Hinton, 2010), etc. Through the use of an appropriate activation function, non-linearities can be introduced in a neural network which enables learning of complex functions.

Deep neural networks are neural networks which have a large number of hid-

den layers[1]. The exact definition is ambiguous, but generally, it refers to neural networks with more than one hidden layer, trained using stochastic gradient descent and its weights updated with backpropagation.

Training A neural network can be trained by providing it with pairs of (x, y) where x is the input and y is the corresponding output. The input is provided to the neural network and the obtained network output $\hat{y}$ is compared to the ground-truth output y. Training neural networks in such a way is referred to as supervised training. Obtaining the optimal solution (the optimal set of weights) for a given problem is an optimization problem. The optimization can be done using stochastic gradient descent and backpropagation. At any given point, the set of weights (the model) is evaluated using an objective function. The objective function is otherwise referred to as a loss or a cost function C. The loss function requires as input the ground-truth output y and the model predicted output $\hat{y}$. The output determines how well the model predicted the output given the input. The loss function is essential in deep learning as its output guides the optimization which in turn produces a desirable solution for the problem.

Estimating how well the model fits the training data can be done using maximum likelihood estimation (MLE). An appealing property of MLE is *consistency*, which states that as the number of training examples goes to infinity the maximum likelihood estimate of a parameter converges to its true value (Goodfellow et al., 2016). A common loss function in multi-class classification problems (machine translation is a multi-class classification problem) and under MLE is cross-entropy which defines the difference between two probability distributions. In essence, it determines how well the predicted distribution matches the empirical distribution.

Word Embeddings Word embeddings are lower dimensional distributed representations of words (Mikolov et al., 2013). Although not relevant to deep learning in general, they are fundamental in NLP. In this section, we discuss word embeddings on a conceptual level. The field of word embeddings is a well-established research area in and on itself, but they are rarely of central importance in MT. Even their common usage in downstream tasks as pretrained word embeddings is seldom employed in MT. Nevertheless, an NMT model uses a word embedding matrix, but learns it from scratch.

In NMT, word embeddings are represented by a trainable matrix where each word is represented with a distributed representation of some dimensionality. When processing a sentence, NMT models map input words to the appropriate represen-

[1]Recurrent neural networks are generally considered "deep" even if having one hidden layer as they are unrolled through time and a single layer is applied multiple times.

tation and use it for all subsequent processing. The embedding matrix is usually randomly initialized.

Although we referred to it as word embedding, commonly NMT models are trained with subword split sentences. For a reasonably sized training dataset, the number of unique words can be very high, which is often prohibitively expensive in terms of memory usage because of the very large embedding matrix. One way to address this issue is to impose a vocabulary threshold and replace all remaining words with a special token. However, this poses a challenge when the special token appears in the output.

The solution to this problem is to use subwords, commonly obtained by applying BPE-splitting (Sennrich et al., 2016b). This is a frequency-based subword splitting which creates a vocabulary of limited size, produces no unseen words and effectively deals with the open vocabulary problem. We use BPE-splitting across the thesis in the training of all models in our experiments.

1.2.2 Recurrent Neural Networks

RNNs are deep neural network models where the output of a given computation depends on the output of previous computations. This naturally lends itself to sequential problems such as time series or natural language. Theoretically, RNNs can model sequences of any length, provided that intermediate states are unnecessary for the task at hand.

The general architecture of an RNN is outlined in Figure 1.1. At any given time step t, the network receives an input x^t. For our purposes, x^t is a word from an input sentence. The network also receives as input the output y^{t-1} of the previous time step $t-1$. The network then performs a computation and outputs y^t. The computation at time t is always dependent on the current input and only the previous output. As previously mentioned, this allows for the modeling of arbitrary long sequences. However, in practice, the propagation of information through a large number of nodes is difficult and poses a significant challenge in RNNs.

RNNs have several other appealing properties. Notably, the size of the model is constant as the number of parameters does not depend on the sequence length. The memory requirements of the model can vary depending on the task. For example, in a single prediction classification task such as sentiment analysis, the full internal representation of the network may not be necessary. More specifically, the intermediate hidden states can be ignored and one can only keep the last hidden state to then use to make a prediction whether the text caries a positive or a

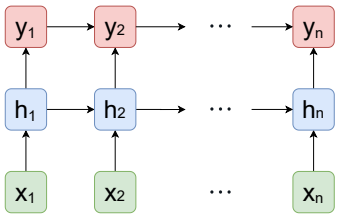


Figure 1.1 – *Recurrent neural network.*

negative sentiment. To some extent, the same holds for machine translation as well. Sutskever et al. (2014) do exactly this, and only keep the last hidden state of the encoder RNN to then condition the decoder RNN and start generating the translation. However, MT is an inherently more challenging task than sentiment analysis. Compressing all the necessary information for an adequate translation in a relatively low-dimensional distributed representation may not be easily achievable.

Despite seeming to be a natural fit to handle NLP tasks, RNNs have disadvantages. They perform all the computations in sequence and are therefore slow. Before being able to perform the computation for time step t, the network has to have already computed the hidden state for time step $t-1$. For large input sequences, this can take a long time. Such architectures are not capable of making use of the high parallelizability of modern hardware. Convolutional neural networks are one example of deep neural networks that can perform computation in parallel which have been extensively used in computer vision. Gehring et al. (2017) have proposed a convolutional sequence-to-sequence model for NMT. However, these models have largely been superseded by the Transformer, which we discuss in Section 1.3.3, and we will omit any further discussion of them.

A long-standing problem with vanilla RNNs is the issue of vanishing and exploding gradients. This issue arises because multiplying gradients across many layers can cause the gradients to explode or alternatively vanish which results in these networks not being easily trainable in a stable way. The problem was largely addressed with the introduction of the Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and its subsequent simplification Gated Recurrent Unit (GRU) (Cho et al., 2014). We first outline the LSTM and later present GRU as it is the variant we use in Chapter 2.

In Figure 1.2, the outline of an LSTM block is presented. As with the vanilla RNN block, the input to a given cell is defined as x_t. The cell receives two additional inputs from the cell at the previous time step. The input shown in the

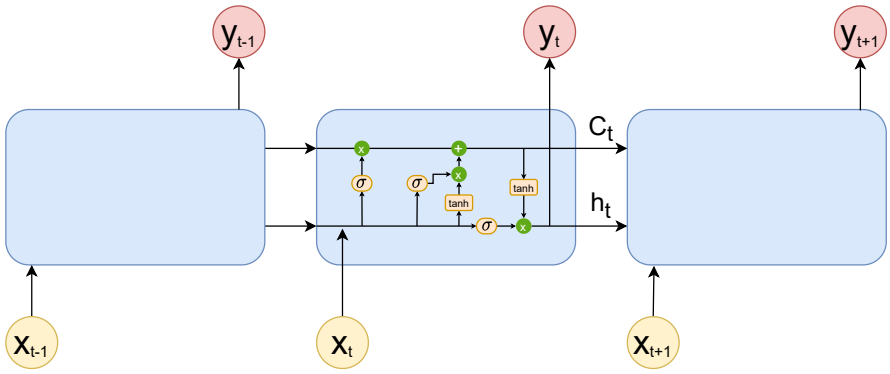


Figure 1.2 – *Long short-term memory recurrent neural network unit.*

bottom left corner is denoted as h_{t-1}. This is combined with x_t by concatenating the two representations and then this resulting representation is passed through a linear transformation:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f).$$

This is also referred to as the *forget gate*. Essentially, the forget gate is in control of how much information of the previous state should be kept for the current computation. Intuitively, this depends on the previous state itself h_{t-1} and the current input x_t.

The next step is modeling the "external" input to the cell. The network performs a similar computation as with the forget gate, to determine the *input gate*:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i).$$

This is the same computation, but it is modeled by a different set of learned parameters W_i and b_i and it is used to control what information of the current cell to update with the input. Subsequently, a candidate cell state $\tilde{C}_t$ is computed by:

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C).$$

At this point, the network has all the necessary information to update its cell state. The cell state C_t is initially set to the previous cell state C_{t-1}. It is then modified by the *forget gate* and updated with the current candidate state:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t.$$

This cell state is later passed through to the cell at the next time step. The actual output of the cell is modified by the *output gate*. The output gate is computed similarly to the other gates:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o).$$

The intuition is that only certain parts of the cell state are relevant for subsequent computations. Unlike the other gates, before the use of the output gate, the cell state is passed through a *tanh* non-linearity. The final output is given as:

$$h_t = o_t * tanh(C_t).$$

The GRU is a simplification of the LSTM. Most importantly, it removes the notion of separate cell and hidden states and it merges them into one and it combines the input and forget gates into a single update gate. The set of equations that govern its internal working are defined as:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \\ \tilde{h_t} &= tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h_t} \end{aligned}$$

Intuitively, one would run an RNN in a forward or left-to-right fashion. This means that the network processes the word at time step $t-1$ before it processes the word at time step t. However, this leads to the words near the end of the sentence being more influential (or being represented more) in the final representations. The words near the beginning of the sentence will not be sufficiently modeled. This can be fixed by using a bidirectional RNN. In addition to the standard forward RNN, one can apply a reverse RNN which processes the sentence from the end to the beginning. The final representation of the RNN can be a computation of the final forward and reverse representation, be it an average, a sum or a concatenation of the two. However, this is still not a fundamental solution to the problem as in long sequences, the words around the middle will not be reasonable represented as both the forward and reverse RNN will fail to adequately incorporate them in the final representation. This drawback is addressed with the introduction of the attention mechanism.

1.2.3 Attention

In this section, we briefly discuss the notion of attention in neural networks and go into more detail in its application in attention-based RNN and Transformer NMT models. Attention in neural networks is loosely inspired by biological attention (with some rather significant implementational differences (Lindsay, 2020)). The key advantage of attention is that it can deal with sequences of arbitrary length and has direct access to individual inputs from the sequence. The utility of attention in NLP is intuitive: instead of summarizing a whole sentence in a fixed representation, as usually done with RNNs, one can use attention to allow for direct access to all RNN hidden states and decide which ones are important for the task at hand. A high-level depiction of attention in an RNN is presented in Figure 1.3.

For classification tasks, such as sentiment analysis, attention can focus on certain words which have been learned to conduct opinion or emotion. For translation, at each time step of the target sentence generation, attention can be used to determine which source sentence hidden representation is important for that particular time step (e.g., when translating a pronoun in German, it is useful to pay attention to the counterpart pronoun on the English source side).

Attention is based on three concepts, queries, keys and values. Although the queries, keys and values can be different, in most cases, the keys and values are the same. In self-attention (which we discuss later), the queries are the same as the keys and values. Conceptually, attention determines the importance of the values by computing scores with which it can weigh them. The scores are obtained from the interaction of the queries and the keys. In NMT, it is common for the encoder hidden states to represent the keys and values and the current decoder hidden state to represent the query. The interaction of the query with each key determines which encoder hidden state is important for the current translation and the obtained attention scores can be used to weigh the values (encoder hidden states).

The method works by computing the so-called attention weights, which determine the importance of each hidden state. Attention weights are usually normalized to sum to 1 by applying softmax. The final result obtained from the attention mechanism is a fixed representation obtained by summing the weighted hidden state representations.

The attention weights can be obtained in several ways. Bahdanau et al. (2015) introduced the additive attention. Formally, it is computed as:

$$attn(q_i, k_j) = v^\top tanh(W[q_i; k_j]).$$

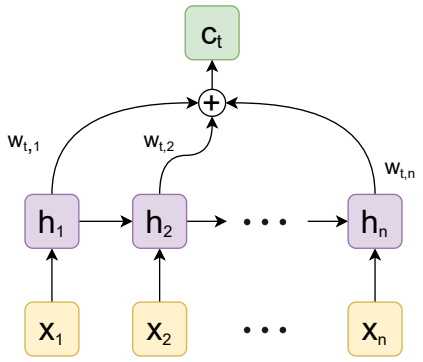


Figure 1.3 – *Attention in a recurrent neural network.*

The drawback of this attention mechanism is the need for optimizing the set of parameters v and W. An alternative to this approach is the dot-product attention (Luong et al., 2015). It requires no additional parameters to optimize and is defined as:

$$attn(q_i, k_j) = q_i^\top k_j.$$

However, the scale of the output depends on the dimensionality of the q and k representations. As a result, Vaswani et al. (2017) introduce the scaled dot-product attention which divides the term by the square root of the dimensionality of the queries and keys. Finally, the attention weights α_{ij} are defined as $\alpha_{ij} = attn(q_i, k_j)$ and a representation s given q_i can be computed as $s = \sum_j \alpha_{ij} v_j$.

1.3 Neural Machine Translation

Neural machine translation is the de facto standard method for training MT systems. This paradigm was introduced with the work of Sutskever et al. (2014) and Bahdanau et al. (2015). Both works are based on a recurrent neural network with the notable distinction that Bahdanau et al. (2015) employ an attention mechanism. Attention has proven to be of fundamental importance for NMT and subsequently to many NLP tasks and beyond language as well.

1.3.1 RNN-based Neural Machine Translation

In this subsection, we will discuss NMT methods based on RNNs. For clarity, we present the attention-based RNN methods in Section 1.3.2. The majority of the work in this thesis is based on the Transformer, but the initial work was partly conducted with the use of RNN-based models.

So far we have discussed a single RNN. In order to enable this network to generate a sequence and in turn do machine translation, we must introduce the encoder-decoder framework. The problem is formally defined as:

$$p(y_1, \dots y_{T'} | x_1, \dots x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots y_{t-1}).$$

The model is tasked with predicting the target sequence $\boldsymbol{y}$ given the input sequence $\boldsymbol{x}$. The length of these sequences can differ as is common in natural language. Sutskever et al. (2014) propose to solve the task by applying an LSTM over the input sequence which results in a single representation v of the input. The generation of the target language translation is conducted by outputting a single token at a time, which is modeled by $p(y_t|v, y_1, \dots y_{t-1})$. The generation of a token y_t is conditioned on the representation of the input v and all the previously outputted tokens y_i where $i < t$. The probability $p(y_t)$ is represented by a softmax over all of the items in the target vocabulary.

Practically, a special end-of-sentence <EOS> token is appended at the end of the source and target sentence. This is necessary to determine when the generation of the translation is finished. As a result, this type of framework can take as input and output unlimited sequences.

A practical consideration which was reported to be useful for ease of training by Sutskever et al. (2014) is the order reversal of the input words. Instead of supplying the input sentence in the ordinary left-to-right manner, it is reversed. In this way, when generating the first target word, the representation of the first source word will be strongly represented in the final input sequence representation. Analogously, for the second word, third word and so on.

It is not uncommon in vanilla RNN-based (LSTM or GRU) encoder-decoder framework to make use of several RNN layers, both in the encoder and in the decoder. Although in a sense, an RNN can be considered as a deep neural network, one can stack several RNN layers on top of each other. This allows for a greater computational capacity and modeling capabilities.

1.3.2 Attention-based Neural Machine Translation

One of the fundamental limitations of a standard encoder-decoder framework is the fact that the source sentence is mapped to a fixed representation. This inherently creates a bottleneck as limited information can be encoded in a representation of a reasonable size. Machine translation is a complex task and it is difficult to compress all the necessary information for a meaningful translation. Bahdanau et al. (2015) introduce the key concept of attention that aims to solve this limitation. Instead of summarizing the full sentence by taking the last hidden state of an RNN, they propose to keep all intermediate hidden states and enable direct access to them in the decoding (generation) process. In order to enable efficient use of all of these hidden states, they propose the attention mechanism. On a high level, attention determines which of the source hidden states are important at any given decoding time step by computing values that score the source hidden states. In essence, the source sequence is still mapped to a single representation, but this representation is different for any decoding time step and depends on what is useful for the generation of the current target word. An overview of the architecture is presented in Figure 1.4.

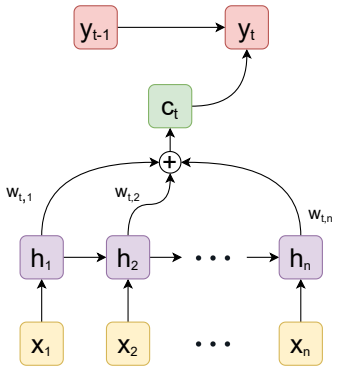


Figure 1.4 – *A neural machine translation model with an attention-based recurrent neural network.*

The model computes a contextual vector c_j at any given decoding time step which is then integrated into the decoder hidden state. The contextual vector is computed as a weighted sum of the encoder hidden states. The following equations formally introduce the model presented in Bahdanau et al. (2015):

$$
\begin{aligned}
c_j &= \sum_{i}^{T_x} \alpha_{ij} h_i \\
\alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{k=1}^{Tx} \exp(e_{kj})} \\
e_{ij} &= v_a^\intercal \tanh\left(U_a s'_j + W_a h_i\right)
\end{aligned}
$$

The weighted sum is based on the attention weights α_{ij} which are the normalized e_{ij} scores. These are computed using additive attention and depend on the current decoder and relevant encoder state.

Even though these models can be used to obtain strong translation quality, they have some limitations. Firstly, they are still fundamentally based on RNNs which are difficult to parallelize. Furthermore, the interaction between arbitrary encoder states is limited. Although the decoder has direct access to all encoder hidden states, when modeling the input sequence, there are limited dependencies between arbitrary source language tokens. These limitations are addressed in the Transformer architecture which we discuss in the following section.

1.3.3 Transformer Neural Machine Translation

Transformer (Vaswani et al., 2017) is an encoder-decoder framework that is entirely based on attention and does not have any notion of recurrence. The Transformer is fundamentally based on the concept of self-attention. Self-attention or intra-attention (Cheng et al., 2016) models the input sequence by relating all individual inputs to all other inputs. This creates a direct dependency of each token to all other tokens and avoids the problems of information propagation faced by RNNs. Furthermore, the modeling of a sentence has no dependencies on previous computations as opposed to RNN-based architectures. Computing the representation for token x_t is independent of the computation for token x_{t-1}, therefore the computation of the full sequence representation can be completely parallelized. This enables the model to train significantly faster than other architectures. The model architecture is shown in Figure 1.5. The Transformer encoder consists of two main subcomponents. The first one is the multi-head self-attention which models the interaction between the inputs (the sentence tokens). The second component is a standard feed-forward neural network which is applied independently and identically on all inputs.

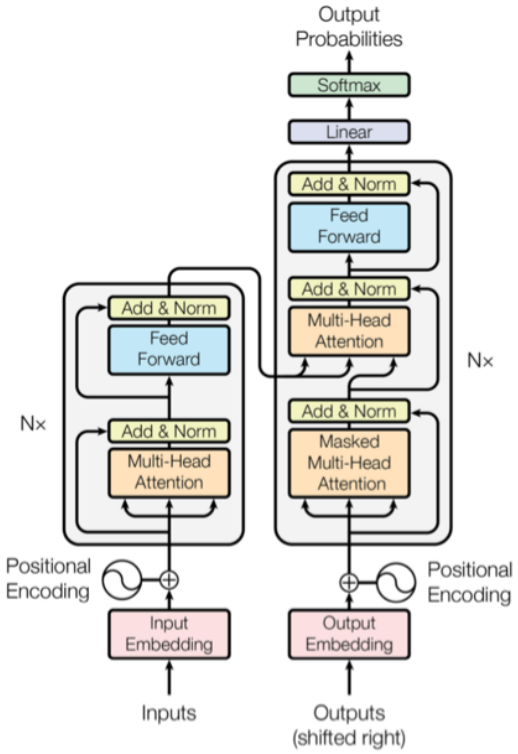


Figure 1.5 – *Transformer encoder-decoder architecture (figure from Vaswani et al. (2017), page 3).*

Similar to all other NLP models, the first step in the Transformer is mapping the inputs to the appropriate input embeddings. Following Press and Wolf (2017), the input, output and linear output layer parameters are usually shared. Press and Wolf (2017) showed that this does not decrease performance and significantly reduces the number of model parameters as the embedding matrices are usually very large for any reasonable number of distinct input tokens (e.g., >30K vocabulary items).

Unlike in RNNs, which have a built-in notion of position arising from the sequential processing of the input, the Transformer has no positional information bias in its architecture. The self-attention mechanism considers the input as a bag-of-words and therefore has no information as to what the order of the inputs is, which is naturally essential for most NLP tasks. The issue is addressed by learning separate positional embeddings. Similarly to learning the word embeddings, the model learns an embedding for position 1, position 2 and so on. Using

this approach, the number of positions must be fixed to a predefined number. If at inference time, the model encounters a sequence of length higher than the predefined maximum number of positions, it will not have a meaningful positional embedding for the remaining tokens. In practice, a sufficiently high number can be set (e.g., 1024 positional embeddings) which will handle the vast majority of possible inputs. Vaswani et al. (2017) propose using sinusoidal positional embeddings which theoretically can extrapolate to unseen positions.

As shown in Figure 1.5, the multi-head attention component of the encoder takes three inputs which are identical. The input to the first layer is the sum of token-level and positional embeddings (subsequent layers take the outputs of previous layers). The three inputs are the queries Q, keys K and values V which are used in the attention mechanism. Q, K and V are the same in self-attention as the goal is to model the interdependencies of the input. The Transformers uses the scaled dot-product attention. The detailed way of how attention works in the Transformer is presented in Figure 1.6.

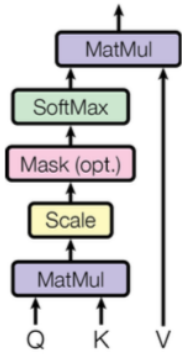


Figure 1.6 – *Attention in Transformer (figure from Vaswani et al. (2017), page 4).*

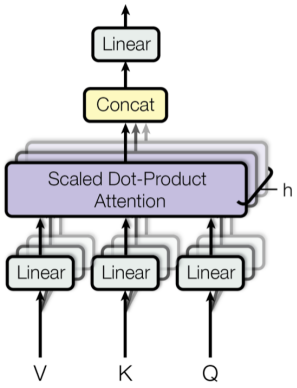


Figure 1.7 – *Multi-head attention in Transformer (figure from Vaswani et al. (2017), page 4).*

Formally, the Transformer attention is computed by:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V$$

The optional masking shown in Figure 1.6 is used in the decoder self-attention. The masking disables access to future positions, meaning that the representation of token y_t depends only on tokens y_i where $i < t$. This is necessary as the model

has access to all target tokens at training time, but does not have it at inference and this condition must be simulated at training time.

We previously discussed that the Q, K and V are the same as the input to the Transformer encoder I. However, in practice, the input is mapped to Q, K and V using separate learnable matrices, W_q, W_k and W_v, respectively. This gives more computational power to the model, but also may be used to expose different properties of the input for the different roles it may have in the attention (different properties of an input token may be useful when it acts as a query compared to when it acts as a key).

Another key concept introduced in the Transformer is multi-head attention. Multi-head attention differs from standard single-head attention in that it applies the attention mechanism multiple times. Generally, this allows for a separation of concerns effect, where different attention heads can learn to model different properties of the input. For example, an attention head can be responsible for modeling any potential coreferential relationships in the input. In Figure 1.7 we depict the multi-head attention in detail.

Considering that multiple attention heads increase the computational requirements of the model, Vaswani et al. (2017) propose projecting the queries, keys and vectors to a lower-dimensional space before applying the separate attention heads. Usually, they are projected to d/h where d is the dimensionality of the Transformer and h is the number of attention heads. This provides for a compromise in computational efficiency.

The output of the multi-head attention is put through two additional operations. First, the model applies the residual connection (He et al., 2016). This operation simply adds the input and output of the multi-head attention. Residual connections have been shown to help train very deep neural networks. A possible reason is the easier passing of gradient in the backpropagation step by providing a simple path with no non-linear functions which may cause gradients to explode or vanish. The second operation is layer normalization which is another way to address the issue of large depth.

The second subcomponent of a Transformer encoder is the feed-forward neural network. The same network is applied to each token of the input independently. This subcomponent has no inductive bias in terms of processing natural language and can be viewed as simply providing computational power to the model.

The Transformer decoder is similar to the encoder. It takes as input the target sentence and it first applies multi-head self-attention with masking. Afterward, it applies attention in a similar manner as to Bahdanau et al. (2015). In this case, the query is the output of the decoder self-attention and the keys and values are

the output of the last encoder layer. This subcomponent enables the model to condition the translation on the input sequence. Finally, the feed-forward network is applied. The outputs of the decoder are projected using a linear layer to the size of the vocabulary and a softmax is run in order to obtain the probabilities for the most likely translation.

1.4 Context-Aware Neural Machine Translation

The majority of academic research and commercial systems work on the sentence-level. Specifically, the systems take as input a single sentence in the source language and output a single sentence in the target language. To the best of our knowledge, most commercial systems, when presented with multiple sentences, first perform sentence splitting and subsequently translate sentences independently. As a result, any potential inter-dependencies between the sentences or the larger document structure are not modeled and are not taken into account in the generation of the translation. This leads to the disregard of any relevant discourse-level phenomena. While a translation of a sentence may appear adequate when looked at in isolation, its validity in terms of document-level coherence may be lacking (Läubli et al., 2018; Toral et al., 2018).

In this thesis, we focus on using document contextual information, meaning the text that is surrounding any given sentence. Despite simplifications made in previous work, text rarely comes in an isolated form and it is usually a part of some document. Modeling context and incorporating any aspects of relevance from it in the translation of a sentence is of high importance.

1.4.1 Related Work

The field of context-aware NMT has sparked great interest in the research community. Aiming at addressing the deficiencies of sentence-level models, numerous works have proposed various ways of using contextual information.

Initial works on context-aware NMT are Jean et al. (2017); Wang et al. (2017); Tu et al. (2017); Tiedemann and Scherrer (2017). Jean et al. (2017) propose an additional encoder responsible for modeling the previous sentence. Wang et al. (2017) train an RNN on the 3 previous sentences and use that representation as contextual information. A key feature of context-aware models is how the contextual information is integrated into the NMT model. Most models follow a similar paradigm, namely, the gating mechanism (Wang et al., 2017). The gating

mechanism assumes a contextual representation is already computed as well as the so-called main (current) sentence representation. The gate controls how much contextual information should be used for the final representation as opposed to the main sentence. Usually this is modeled by $f = z * m + (1 - z) * c$ where m is the main sentence representation, c is the context representation and z is the computed gate which is usually conditioned on m and c.

Tu et al. (2017) propose a method that continuously builds a cache storing certain decoder states which can then be accessed in translation. Tong et al. (2020) is a similar work where they cache encoder states. The advantage of this model is that it does not need to compute the contextual representation from scratch for each sentence, but it has limited modeling capabilities as it is very difficult to model certain discourse phenomena, such as anaphora resolution. Even if we assume that the decoder representations of all the possible antecedents of an anaphoric pronoun are cached, it is difficult to determine which is the correct antecedent. The translation model views the cached representations in isolation from the text in which they originally appeared. Therefore, reasoning over the cache to determine the correct antecedent is practically very difficult.

Tiedemann and Scherrer (2017) propose concatenating consecutive sentences and applying a standard NMT model. They propose prepending the previous sentence and inserting a special sentence-separating token between the two sentences. The concatenation can be done on the input side or on the output side as well. Naturally, this method can be extended to include an arbitrary number of contextual sentences, but is limited by memory requirements. Tiedemann and Scherrer (2017) use RNN-based NMT. In Chapter 2 and Stojanovski and Fraser (2018) we present results using a Transformer concatenation model. Ma et al. (2020) propose a slight modification to this approach by ignoring context in the upper layers of the model. Zhang et al. (2020) introduce a separate main sentence attention and combine it with the global attention that uses the context as well.

Voita et al. (2018); Miculicich et al. (2018); Zhang et al. (2018a); Stojanovski and Fraser (2018) proposed the initial context-aware Transformer models. Voita et al. (2018) propose a separate context encoder only in the last layer of the Transformer. The main sentence and context encoder representations are merged using a gating mechanism. In Chapter 2 and 3, we show work where we proposed similar models, but the contextual information is integrated directly into the decoder. Zhang et al. (2018a) is a similar work as well where they used the contextual information in the encoder and decoder and show the effect of the different ways of integrating the context. Miculicich et al. (2018) proposed a more elaborate model that uses larger context where first sentence-level attention is computed

and subsequently a token-level attention. Tan et al. (2019) propose a hierarchical architecture as well where sentence-level representations are used to compute a document-level representation which is then used in the decoder.

Maruf and Haffari (2018) use a memory network to store source and target context and use attention over them to determine their importance at each decoding step. The source memory is represented as the hidden state of a document-level bidirectional RNN which is applied on the sentence-level representations of the document. The target memory is represented with the last decoder hidden state of each context sentence.

Maruf et al. (2019a) propose using selective attention in order to be able to learn across long distances. The method works by computing sentence-level attention first which is similar to the method of Miculicich et al. (2018). However, in order to make it scalable across realistic document lengths (e.g., longer than 3 sentences) they employ sparsemax (Martins and Astudillo, 2016) which cuts of gradient backpropagation through attention links which have very low attention scores by setting it to zero. After this operation, the method performs word-level attention which is likely scalable at this point as it performs attention over a small number of sentences. Yang et al. (2019a) propose a query-guided capsule network to include contextual information. The model uses a dynamic routing algorithm to retrieve relevant contextual features from the previous sentences.

One of the key challenges in context-aware NMT is the accurate measurement of improvements with regard to discourse. Hardmeier (2012) point out that the standard MT quality metric BLEU (Papineni et al., 2002) cannot be reliably used to determine better handling of discourse-level phenomena. Several works attempt to address this issue by proposing various methods of evaluation. Bawden et al. (2018) propose several RNN-based context-aware NMT models, but more importantly, focus on evaluation in context-aware NMT by manually creating challenge sets designed specifically for evaluating coreference resolution, coherence and cohesion. Müller et al. (2018) propose a challenge test set called ContraPro for evaluating coreference resolution which is automatically created and because of the significantly higher number of test sentences, is more robust than small scale manually created test sets. In this thesis, we make use of this work, particularly in Chapter 3 where we use it as one of the metrics to show the effectiveness of our approach and in Chapter 6 where we question to what extent should accuracy on ContraPro alone be used to make claims that coreference resolution is being solved by a given NMT model.

Voita et al. (2019b) propose challenge sets tailored to specific discourse-level phenomena for English→Russian translation. Furthermore, they propose a context-

aware model that works in two stages, first individual sentences are translated by a context-agnostic model and are later refined by a context-aware model. The context-aware model uses the sentence-level model's encoder states. Voita et al. (2019a) modify this approach by only using the output of the sentence-level model and not any model states, thus enabling this approach to easily work for any model architecture. Both models show large improvements on discourse-specific challenge sets, but do not provide for significant improvements in terms of BLEU scores.

Several works use reinforcement learning for context-aware NMT. Kang et al. (2020) propose using context selection algorithms to determine useful sentences. This approach enables using fewer and potentially more informative context sentences which contributes to more efficient training and generation, and better performance. However, this approach may lead to worse anaphoric pronoun translation as context sentences containing the corresponding antecedents may be filtered out. Saunders et al. (2020) optimize document BLEU with Minimum Risk Training. Jauregi Unanue et al. (2020) use discourse rewards to provide for better lexical cohesion and coherence in MT.

Kim et al. (2019) conduct a detailed manual analysis of what kind of errors do context-aware NMT models fix in relation to sentence-level models. While they observe some cases where the translation fixes are related to coreference and topic-aware lexical choice, the majority of improvements are not interpretable. They make the case that it is likely that document-level models provide a regularization signal and that in order to realistically estimate any improvements from them, they must be compared against strong sentence-level models. Similar conclusions are made in Li et al. (2020). In all our works presented in this thesis, we compared against strong sentence-level models. We always train 6-layer Transformer models with a hidden size of at least 512 (in Chapter 5, we use a hidden size of 1024) and train models on significant amounts of parallel data (at least 4.5M up to 22M training examples). We hypothesize that the larger improvements we have seen in our works are because: (1) we used OpenSubtitles which is a domain very dependent on context; (2) in multi-domain settings coherence is of pronounced importance as domain-dependent decisions are more prevalent as opposed to a single-domain setup.

1.4.2 Context-Aware NMT Model Taxonomy

In the following, we will discuss previous works and attempt to provide a taxonomy of context-aware NMT models. This is not a strict taxonomy and is outlined

to provide for an easier understanding of the different context-aware models. This is aimed at helping to provide a clear distinction of different aspects of context-aware models.

Input-Output

To a large extent, previous work has used the term "context-aware NMT" to refer to models that use contextual information on the input side, but ones which are essentially sentence-level models on the output side. Specifically, these models have a notion of a main or current sentence which is to be translated and the context is used as auxiliary information. Most commonly, the context representation is recomputed when translating the subsequent sentence. We call these models input-based context-aware models. An alternative is to output more than a single sentence which we call output-based models. We do not differentiate between using source or target contextual information as input. The majority of previous works fall under the category of input-based context-aware models (Jean et al., 2017; Wang et al., 2017; Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Tu et al., 2017; Zhang et al., 2018a; Maruf and Haffari, 2018; Maruf et al., 2019a; Stojanovski and Fraser, 2019a,b, 2020; Voita et al., 2019a,b, *inter alia*). These models receive the source sentence as input as well as contextual information.

We also classify the work of Tiedemann and Scherrer (2017) under this category. Although the type of model they propose can be used to produce more than a single-sentence translation, this is not how the model is used in Tiedemann and Scherrer (2017). The output of the model is split based on a special sentence-separating token and the translation after this token is considered as the main translation.

Output-based context-aware models output context and use it as the final translation. Junczys-Dowmunt (2019) follow the same approach as Tiedemann and Scherrer (2017) with the key difference that they consider large paragraphs as input (up to 1000 tokens) and use the output of this model as the full translation of the input. A similar approach is taken in the work of Liu et al. (2020). Taking this approach requires some care when evaluating the models. BLEU is usually computed by comparing individual sentences. This can be addressed by splitting the document into sentences based on special sentence separating tokens (Junczys-Dowmunt, 2019) or computing document-level BLEU (Liu et al., 2020).

Cache-based models (Tu et al., 2017) fall on the boundary of this classification. These models output a single translation at a time, but do not recompute the

context representation for each sentence and continuously update this representation as they translate a document. As a result, one can take the perspective that these models translate a document instead of an individual sentence. Despite this, we classify these models as input-based context-aware models as the interaction between the output translations is limited, unlike the approach taken in Junczys-Dowmunt (2019); Liu et al. (2020).

Source-Target

The other distinguishing feature of context-aware models is whether they use source or target contextual sentences. Using source context sentences allows for faster models as the translation of the current sentence does not depend on the translation of the previous sentences. This applies only at inference time. These models may be of more practical use as a given input document can be split into individual sentences which can be translated in parallel while still using contextual information. However, the target sentences contain more information which is helpful in translation. For example, if a model has access to the information that "bank" was translated as a "financial bank" in a previous sentence, it will be more certain that the "bank" in the current sentence should not be translated as a "river bank".

Apart from the computational efficiency issue which we discussed, these models suffer from the *exposure bias* phenomenon. Exposure bias is usually used to refer to the problem that standard NMT models face, namely that in training, the model only sees gold-standard previous translation decisions and is never exposed to its own potential translation mistakes. At inference time, the model does not have access to gold-standard translation decisions and can therefore easily find itself in an out-of-distribution situation. The same problem is present in context-aware models that use target context. At training, the reference target context sentences are used, but at inference, the model needs to use machine-translated context. As opposed to exposure bias in standard NMT, in context-aware models, the issue can be easily addressed by using machine-translated target context during training. Mino et al. (2020) address this issue and propose alternating between using reference and translated target context sentences. Finally, a model can use both source and target context.

We make the case that what type of context to use depends on a case-by-case basis. If translation speed is of the essence, target context should not be used. Alternatively, if translation performance is more important, target context should be more helpful in providing more coherent translations. However, we are not

aware of conclusive experiments showing this to be the case.

Previous-Subsequent

Context-aware models also differ based on whether they use previous sentences as context or subsequent ones as well. Maruf and Haffari (2018) make this distinction as well and use the terms online or offline modes. Previous context models can use both source and target context sentences. Subsequent context models usually only use source context as in principle, the future target sentences are not available. However, two-pass models (Xiong et al., 2019) where candidate translations are first generated and are subsequently used for the final translation can be considered as models that use future target context.

While the utility of subsequent context sentences may not be immediately obvious, they can contain useful information for word-sense disambiguation or cataphora resolution. Wong et al. (2020) study the translation of cataphoric pronouns, but also show that they are considerably less frequent than anaphoric pronouns. Therefore, in this thesis, we focus solely on anaphoric pronoun translation and the use of previous context sentences.

Local vs Global Context

Intuitively, this distinction differs models by whether they use local context or take into account the full document. The definition of local context is fluid, but one can make an arbitrary threshold of a small number of sentences which we believe can be reasonably set to 3 sentences.

A number of works (Jean et al., 2017; Tiedemann and Scherrer, 2017; Voita et al., 2018, *inter alia*) have been proposed that make use of local context. The experimental evidence seems to suggest that local context is being utilized in a meaningful way and improvements are being obtained which are measurable on discourse-specific metrics. Several works have proposed models that are capable of taking into account the global context (Maruf et al., 2019a; Junczys-Dowmunt, 2019, *inter alia*). In addition to their computational efficiency limitations, to the best of our knowledge, there is little evidence that using global context provides for reliable improvements. While these models improve over sentence-level baselines, it is unclear whether modeling global context is a significant contributing factor.

In contrast to NMT, language models using long-distance context have unambiguously proved to be very effective (Dai et al., 2019; Yang et al., 2019b; Rae et al., 2020; Kitaev et al., 2020; Beltagy et al., 2020). Our assumption is that

global context is important for translation as well. However, there are two major aspects that may hinder the role of global context in NMT. Firstly, more work is necessary to properly evaluate the utility of global context compared to local context. The majority of evaluation metrics in MT do not attempt to discern the contribution of these two types of context in the translation performance. Secondly, the signal from global context may be quite faint for the translation of a single sentence compared to the information coming from the current source sentence. Therefore, we assume that more work is necessary to design context-aware NMT models capable of identifying useful global contextual information.

Fine- and coarse-grained

We previously discussed that a large number of previous works in context-aware NMT use a similar technique for integrating context, namely the gating mechanism. However, how they model the context differs. While this is true, the majority of context-aware architectures model the context, in some sense, in a fine-grained way. Specifically, when integrating the contextual information in the standard NMT architecture, the model is given fine-grained access to the context, usually being able to access the token-level representation of the context. If the representation given to the NMT model is to some extent abstracted away from the token level, it is usually built by low-level complex interaction of the tokens, often using self-attention. While this gives the models larger modeling capabilities, the computations involved to obtain such representations are very expensive (consider translating an entire book).

We make the case in Chapter 4 and 5 that a fine-grained representation is not always necessary and that coarse-grained modeling of context can be sufficient in some cases. In Chapter 5 we continue along these lines and make the case that local context should be modeled in a fine-grained way while global context in a coarse-grained way. For example, anaphora resolution depends on fairly local contextual information which needs to be modeled in a fine-grained way in order to access information about the potential antecedents. However, choosing the correct word sense (financial bank as opposed to a river bank) or the correct formality (T-V distinction, related to deixis) may be inferred from relatively global information (e.g., domain) which can be represented in a more abstract way, which in turn implies that coarse-grained modeling may be sufficient. To our knowledge, except for the work in this thesis, no other work has attempted to make this distinction in a clear and concise way. Imposing this either as an explicit architectural decision (as we do in Chapter 5) or implicit bias in context-aware models is a promising

future direction.

1.5 Discourse in Machine Translation

Sentence-level NMT has achieved very high-quality translation across many language pairs. Given certain conditions, such as availability of very large parallel data and no domain shift, some previous works have even claimed to achieve human parity, notably Hassan et al. (2018) for Chinese→English news translation. In Hassan et al. (2018), this is ascertained by asking human evaluators to score translations using direct assessment and later these scores are used to determine whether human and machine translation scores are statistically indistinguishable. However, this claim has been challenged in the works of Läubli et al. (2018); Toral et al. (2018), scrutinizing several aspects of the manual evaluation conducted in Hassan et al. (2018). The aspect of interest to this thesis is that the human judges evaluated the sentences in a context-agnostic way, meaning that the evaluators were presented with the single source and target language sentence. Läubli et al. (2018) show that when presented with a sentence and its surrounding context, human evaluators still prefer human translations as opposed to machine translations obtained from a sentence-level model. This is intuitive since sentences are rarely standalone and are usually contextualized in some way. In some cases, they can be translated with perfect accuracy without taking context into consideration. However, many properties of human language are dependent on context, or more precisely on discourse. Discourse has long been of interest in the MT community. The DiscoMT workshop (Webber et al., 2013, 2015, 2017; Popescu-Belis et al., 2019) has been conducted several times which enabled focused work on this important problem. Hardmeier (2012) provides a detailed overview of discourse in MT for SMT, while Maruf et al. (2019b) is a more recent survey including NMT. In the following section, we discuss discourse-level phenomena which are of interest for machine translation and ways of evaluating how well are they handled by an MT model.

1.5.1 Discourse-level Phenomena

We focus our analysis on three discourse-level phenomena: coreference (anaphora) resolution, coherence and cohesion. These three phenomena are of key interest in the remainder of the thesis. However, we provide a short overview of other closely related phenomena such as deixis, ellipsis, discourse connectives and other.

Coreference (anaphora) resolution

Coreference resolution is an NLP task which attempts to solve the problem of linking two or more expressions referring to the same entity. An entity can be a person, a non-human animal, an inanimate object or an event. In this thesis, we consider anaphora resolution where the referring expression makes mention of an entity that appeared previously in the text under consideration. An expression can refer to an entity succeeding it, a phenomenon known as cataphora. However, Wong et al. (2020) show that the phenomenon appears considerably more rarely in comparison to anaphora.

A common instance of AR is pronominal anaphora resolution, where the referring expression is a pronoun. Pronominal AR is of interest for MT primarily because of gender.

Take for example, translation from English to German. In English, the third person singular "it" can refer to any antecedent non-human animal or inanimate object because nouns are genderless in English. However, in German, nouns have gender. Therefore, "it" can be translated into "er" (masculine), "sie" (feminine) or "es" (neuter), the translation depending on the German gender of the correct translation of the English antecedent. A challenge appears in the reverse translation direction as well. In German, the masculine "er" or feminine "sie" third person pronoun can refer to a human or object with the corresponding gender. In English however, one needs to make a distinction whether the pronoun refers to a person (in which case "he" or "she" is the appropriate translation) or an object (where "it" is necessary). A challenging problem is when the source and target language are both gendered with regard to nouns. Naturally, noun gender differs across languages. Take for example German and any other language with three gendered pronouns. If the German "er" refers to an inanimate antecedent, it can be translated into any of the three gendered pronouns in the target language. However, statistically, "er" will most commonly be translated into the masculine pronoun and learning to translate otherwise is likely going to be difficult for current MT models.

Finally, many languages are pro-drop languages, meaning that pronouns can be dropped without loss of grammaticality if they can be otherwise unambiguously inferred. Examples include Chinese, Japanese, Korean, Slavic languages and others. Naturally, a challenge arises when translating out of a pro-drop language to a non-pro-drop language where additional inferences need to be made to determine the appropriate pronoun in the target language.

Lapshinova-Koltunski and Hardmeier (2017); Šoštarić et al. (2018) analyze discourse structure alignment discrepancies which lead to missing discourse in-

formation in MT training data. Apart from issues arising from alignment, such discrepancies also originate from contrasts between two languages, where certain structures are subject to epxlicitation or implicitation (Becher, 2011). While they find that NMT models are relatively better at handling such discrepancies, performance still lacks human translation. For English→German they find that many pronouns do not require explicit pronoun translation and observe that naive pronoun models tend to overproduce pronouns on the target. This is a crucial consideration which is relatively overlooked in current context-aware NMT work on anaphora resolution, where the majority of evaluation methods assume an explicit pronoun on the target side.

CR is a challenging problem as often requires world knowledge and strong reasoning, both lying outside the capabilities of current deep learning models. However, training deep learning models that can appear to be solving some rudimentary version of CR is within current capabilities. In Chapter 6, we show that NMT can improve pronoun translation in cases where it depends on CR. However, we conclude that the models rely on brittle heuristics and do not solve CR in a fundamental way.

Cohesion

Cohesion is a discourse phenomenon related to the surface properties of text and the "relations of meaning that exist within the text" (Halliday and Hasan, 1976). One type of cohesion is grammatical cohesion which is based on structural content. Of particular interest in MT has been the second type of cohesion, lexical cohesion. Often in MT, lexical cohesion is seen in the context of repetition, namely, multiple instances of the same source word (or synonyms of it) should be translated with the same target word (or synonyms of it). For example, given two consecutive sentences "Someone told me the bank is far away." and "Actually, he was wrong, the bank is much closer", "bank" should be translated consistently in both cases to "Ufer" or "Bank" or corresponding synonyms of them. Another form of lexical cohesion is collocation which refers to words appearing together more frequently than expected by chance. Phraseological collocation is a notable example of collocation. Halliday and Hasan (1976) show the example of "strong tea" and "powerful tea" being equivalent in meaning, but the former is largely preferred by native speakers. While this phenomenon is of interest, it is less studied in context-aware MT because it is challenging to evaluate.

Coherence

De Beaugrande et al. (1981) define coherence as relating to the consistency of the text to concepts and world knowledge. In essence, coherence relates to how intelligible or semantically meaningful some text is. As a result, it encompasses

several aspects of discourse, such as cohesion (a text needs to be grammatical and lexically meaningful), entity-based coherence and discourse connections between utterances. While coherence is very important for MT, so far it has mostly been addressed by focusing on lexical cohesion. The main reason is the simplicity with which lexical cohesion can be modeled and evaluated. In contrast, evaluating the coherence of a full document is challenging. As pointed out in Blakemore (2002), a text can be grammatical and exhibit a reasonable level of lexical cohesion, but be completely incoherent. Sim Smith (2018) argue that coherence is a cognitive process. As such, current NMT models are unlikely to produce highly coherent text in more challenging cases. Promising results have been shown with language models (Radford et al., 2019) capable of generating seemingly very coherent long text. However, the evidence is anecdotal, evaluated using perplexity, retrieval-based metrics (Cho et al., 2019) or metrics such as the ones proposed in Lapata and Barzilay (2005). Coherence remains a challenging problem, both in terms of modeling and evaluation.

Other aspects of discourse

Deixis relates to the use of words and phrases that refer to a specific time ("now", "then" etc.), location in discourse ("this", "that" etc.) or person. Personal deixis is of interest in MT, particularly when translating from a language that does not have a T-V distinction into one that does. T-V distinction refers to the distinction between formal and informal "you" ("du" and "Sie" in German). This has been studied in Voita et al. (2019b) for English→Russian and Sennrich et al. (2016a) for English→German, although not with regards to context. This problem has so far been relatively understudied in context-aware NMT. In Chapter 4, we conducted a manual analysis and found our simple models do not improve the handling of formality. However, further and more systematic work is necessary to fully understand the issue. An important consideration is that in some cases, the output formality is in fact a user preference and any model inferences about the correct formality based on the input should be overridden accordingly.

Ellipsis is the omission of a clause that can rather be inferred from context. This phenomenon is of interest when translating between languages that do not share a certain specific instance of an ellipsis. In Voita et al. (2019b) they show the example of "Veronica, thank you, but you saw what happened. We all did." whose translation into Russian must contain the verb even in the second sentence.

Discourse connectives or markers are words or phrases that help manage the flow and structure of the discourse. Some examples in English are interpersonal ("look", "exactly" etc.), structural ("first of all") or referential ("now", "because") etc. They can be explicitly expressed or be implicit depending on the language.

Translating between languages with different ways of expressing discourse connectives is therefore not straightforward. In many cases, discourse relations across sentences have to be considered in order to correctly translate discourse connectives. This issue is explored in Meyer et al. (2012) where they use sense labels to improve discourse connectives in SMT. Discourse connectives have been relatively understudied in NMT, with the exception of Cai and Xiong (2020) who propose a test set for evaluating the handling of discourse connectives in English→Chinese translation.

Domain is not considered as a discourse phenomenon, but as a concept, it encompasses several discourse phenomena. Discourse markers are often domain-dependent. Furthermore, providing proper translations within a domain requires appropriate stylistic choices and more prominently, coherent choice of words. This is closely related to lexical cohesion, but unlike the applications of lexical cohesion in MT, in domain, the repetition aspect can be disregarded. The choice of translating "bank" as "river bank" or "financial bank" can often be inferred even if no previous instances of it appear in the text. In Chapter 4, we explore domain with regards to context-aware NMT.

1.5.2 Evaluation

The majority of works in MT evaluate with BLEU as an automatic metric because of its ease of use and to some extent, applicability across many languages. While some works have shown that BLEU correlates well to human judgments (Bojar et al., 2016; Reiter, 2018), Mathur et al. (2020) show that it is in fact less reliable than other metrics. More important for context-aware NMT, Hardmeier (2012) show that BLEU does not capture improvements in modeling discourse-level phenomena. As a result, there have been a number of works that aim to address this issue and propose metrics for evaluating how well an MT model handles specific discourse-level phenomena. We focus on metrics for pronominal anaphora resolution and cohesion which we used in the works in this thesis.

Coreference (anaphora) resolution In a general test set, improvements in pronoun translation due to improved coreference resolution are unlikely to be visible in BLEU. One of the reasons is that such pronouns that can be translated only by doing CR, are not very frequent. Voita et al. (2018) circumvent this issue by creating a test set where each sentence has an anaphoric pronoun. BLEU scores on such a test set are more likely to reflect improvements on anaphora resolution. However, identifying such test sentences may be challenging as the number of available candidates may depend on the domain.

Several works have been proposed that use F_1, partial credit and oracle-guided approaches. Hardmeier and Federico (2010) use precision, recall and F_1 to measure pronoun translation. In order to compute these metrics, they first word align the source to the reference and system translations. Then they determine how often do the pronoun translations match between the reference and system outputs, but also clip the counts by the number of occurrences in the reference. Miculicich Werlen and Popescu-Belis (2017) extend this approach by using heuristics in the determination of whether reference and system pronouns match. Furthermore, they take into account whether certain pronouns are missing in the reference or system output and whether the pronoun translations are identical, equivalent, or incompatible. Guillou and Hardmeier (2016) propose a manually created test set which evaluates by comparing the reference and system output. However, they separate the test set into individual pronoun groups in order to test specific aspects of pronoun translation.

All of these approaches measure pronoun translation in an automatic way. However, Guillou and Hardmeier (2018) show that although fully automatic metrics have some correlation to human judgments, they are not on par with semi-automatic ones. Guillou and Hardmeier (2018) specifically investigate AutoPRF (Hardmeier and Federico, 2010) and APT (Miculicich Werlen and Popescu-Belis, 2017) and show that automatic metrics handle certain linguistic patterns well, but do not provide wide coverage of different pronoun functions.

A different approach to automatic pronoun translation evaluation is the scoring-based method. Sennrich (2017) show that an NMT model's scores can be used for evaluation. They create contrastive translation pairs where one translation is correct and the remaining ones introduce some errors. Subsequently, the model scores can be used in order to determine whether it prefers the correct translation over the erroneous ones. The accuracy of a given model is calculated as how often does it score the correct translation higher. In Sennrich (2017) this approach is used to judge the grammaticality of a sentence, but the approach is easily applicable to other properties of translation such as how well are pronouns translated.

Bawden et al. (2018) propose this approach for evaluating coreference resolution, coherence and cohesion for English→French translation. They manually create contrastive translation pairs, but also create semi-correct translations where the source and target context sentence are not translated completely correctly. However, manually creating challenge sets is laborious and limits the size of the test sets which in turn may not provide for a robust estimation of a model's ability to do coreference resolution. Müller et al. (2018) address this issue for English→German translation by automating the procedure and creating a large

contrastive challenge set of 12,000 examples named ContraPro. A contrastive pair consists of the main source sentence and three translation options. An example is provided in Table 1.1. The main sentence itself is not informative enough to make an unambiguous decision regarding the translation of "it". Each of the three possible translations, "er" (masculine), "sie" (feminine) or "es" (neuter) are possible. However, from the context it is obvious that 'it" refers to "novel" which in this case it is translated to "Roman" in German. The gender of "Roman" in German is masculine which shows that "er" is the correct translation in this case. In Chapter 6, we show that this challenge set is not robust to adversarial attacks and that ContraPro scores can be manipulated by small and unimportant changes to the context sentences.

Source context	Let me summarize the novel for you.
Target context	Ich fasse den Roman[masculine] für dich zusammen.
Source	It presents a problem.
Reference	Er präsentiert ein Problem.
Contrastive 1	Sie präsentiert ein Problem.
Contrastive 2	Es präsentiert ein Problem.

Table 1.1 – *Contrastive translation pair for English→German anaphora resolution.*

Jwalapuram et al. (2019) propose training a separate model for pronoun translation evaluation. The model is trained on pairs of sentences containing the reference and a system translation where the pronouns differ. At evaluation time, the model can be used to determine whether an MT model's translation produced a correct pronoun translation by comparing it with the reference. This approach has an advantage over the scoring-based methods because it evaluates actual model outputs. While the scoring-based approaches test the ability to do coreference resolution, there is no guarantee that this will translate to better pronoun translation when the model is to freely translate a source sentence. For example, a model may prefer the reference translation in Table 1.1, but to our knowledge, there is no conclusive evidence that it would produce "er" if it is to translate the source sentence from scratch. However, the approach in Jwalapuram et al. (2019) still does not address the concern outlined in Guillou and Hardmeier (2018) that often there is no one-to-one correspondence between pronouns in different languages. A translation may be valid even if it does not contain the exact reference pronoun

and it is often determined by how the antecedent was translated. An English word may be equally well translated into two German distinct words of a different gender. For example, "engine" in many contexts can be translated equally correctly into "Motor" (masculine) or "Maschine" (feminine). In this case, evaluating the translation of an "it" referring to "engine" does not only depend on the reference translation, but also on what was chosen as the translation of "engine" in the previous sentence.

Evaluating pronoun translation is challenging. There are several different approaches, each with its advantages and disadvantages. While completely automating the evaluation and simultaneously having a wide coverage of different linguistic patterns may be difficult, testing a model with several different evaluation techniques may provide for a more robust estimation of its capabilities.

Coherence and cohesion These discourse phenomena can also be evaluated with challenge sets. Bawden et al. (2018) provide a test set covering both phenomena. In all examples of the test set, there is a word which does not have an unambiguous translation in the current main sentence. In one subset of the test set, they evaluate alignment, meaning that a source word (e.g., "engine") can be translated in two synonym words (e.g., "Motor" or "Maschine"), but only one of them is valid (say "Motor") because it was used in the previous sentence. They also test cases where the translation is determined by the general semantics of the context, e.g., the context sentence "It is 50$." determines that "steep" should be translated to "happig" (expensive) instead of "steil" (sharply sloped). Lexical cohesion is also evaluated with challenge sets in Voita et al. (2019b).

Wong and Kit (2012) propose to evaluate lexical cohesion as the "ratio between the number of repeated and lexically similar content words over the total number of content words in a document" (Miculicich et al., 2018). They determine the lexical similarity using WordNet (Miller, 1995). Miculicich et al. (2018) also use a metric proposed by Foltz et al. (1998) based on Latent Semantic Analysis (LSA). The metric score for a given document is defined as the average cosine similarity between each two consecutive sentences. The cosine similarity is computed on the LSA representation of the two sentences.

Gong et al. (2015) modify BLEU with a cohesion and a gist consistency score. The gist consistency score is defined as a topic consistency and is measured as the Kullback-Leibler divergence between the reference and system translation represented with LDA. The cohesion score is based on simplified lexical chains.

Apart from these three discourse-level phenomena, there are works focusing on evaluating discourse connectives (Meyer et al., 2012; Smith and Specia, 2018; Cai and Xiong, 2020). Elaborate overview of context-aware models and evalua-

tion is also provided in Maruf et al. (2019b); Popescu-Belis (2019).

1.6 Summary and Overview

In this chapter, we presented the definition of the problem that is addressed in this work as well as an overview of the key concepts that are used in the thesis. We defined machine translation and some key neural architectures that are commonly used and we presented the importance of modeling discourse-level phenomena in translation and how can this be addressed with context-aware neural machine translation. The remaining five chapters are published works related to this problem. Chapter 2 shows a method that can gauge the importance of some key discourse phenomena in MT, namely coreference resolution and coherence, using oracle information. We show that both phenomena are under-modeled in current MT systems. In Chapter 3, we follow up on the work in Chapter 2 and use oracle information to create a curriculum that enables easier modeling of coreference resolution. We show that the method is useful under certain circumstances and we provide some insights on how to best train context-aware NMT models. In Chapter 4, we conduct an investigation of the usefulness of context in modeling domain and test two proposed models in a multi-domain scenario and on zero-resource domains which are not seen in training. We show that the proposed context-aware models improve in this experimental setup. In this chapter, we further show that modeling domain can be efficiently done with simple architectures. In Chapter 5, we build on this intuition and show that modeling global and local context separately can provide for improved performance. In Chapter 6, we present a study that tackles the question of whether coreference resolution is modeled in a meaningful way. We identify weaknesses with current evaluation test sets for coreference resolution in machine translation and propose a new template test that evaluates specific steps of a coreference resolution pipeline. Finally, we propose a simple training data augmentation that improves on pronoun translation as measured by existing challenge sets, but does not fundamentally improve coreference resolution in MT. Our work calls into question the intuition that this problem is easily modeled by current context-aware machine translation models.

Chapter 2

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

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

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.

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.

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.

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

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.

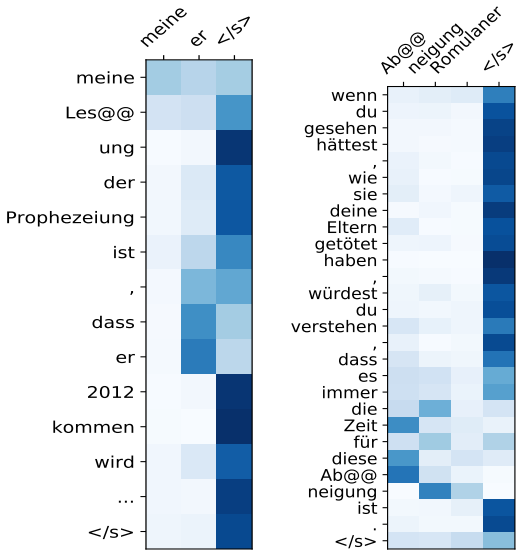
tains small gains for coherence and larger ones for coreference. The context-aware model we trained with the previous target sentence as context, scored 63.0% and 54.0%, on the coreference and coherence challenge set, respectively. From these results we also can conclude that our model is reasonably 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.

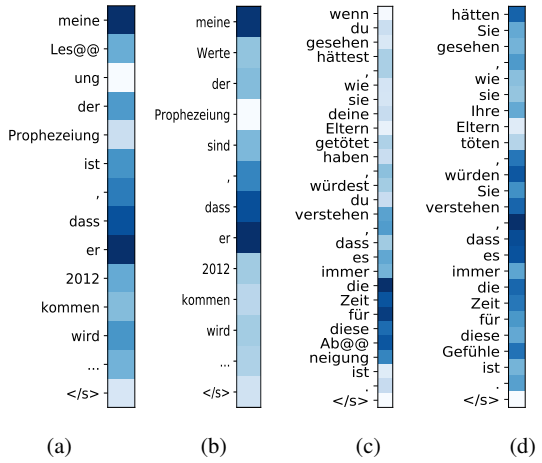


Figure 2: Gate activations for pronoun and repeated words oracles. (a) pronoun oracle, (b) - noisy pronoun oracle, (c) - repeated words oracle, (d) - noisy repeated words oracle.

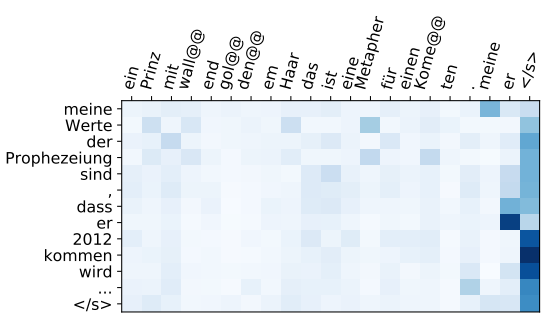


Figure 3: Context attention of our proposed model on the noisy pronoun oracle.

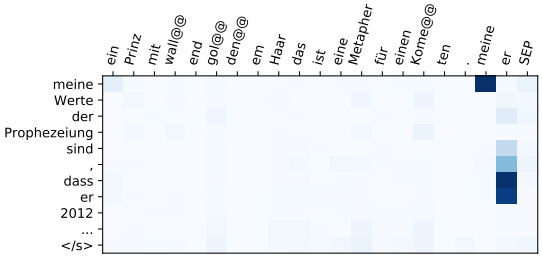


Figure 4: Attention over the previous sentence of the concatenation model on the noisy pronoun oracle.

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*.

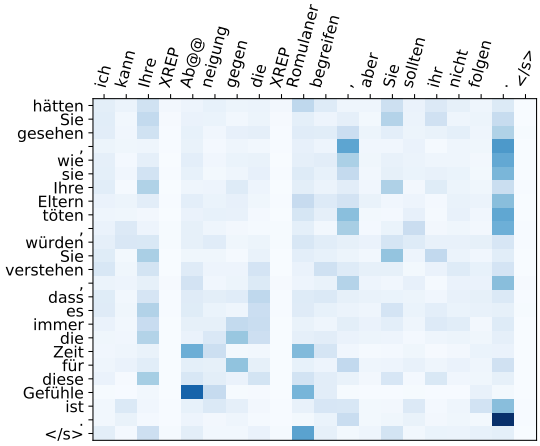


Figure 5: Context attention of our proposed model on the noisy repeated words oracle.

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

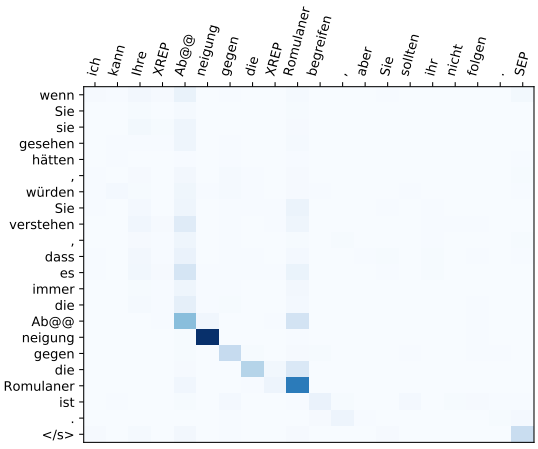


Figure 6: Attention over the previous sentence of the concatenation model on the noisy repeated words oracle.

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.

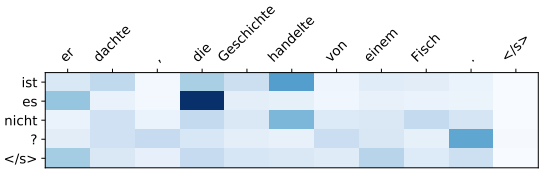


Figure 7: Context attention of our proposed model on the previous target sentence.

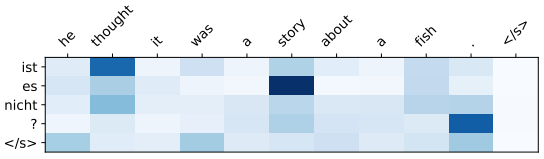


Figure 8: Context attention of our proposed model on the previous source sentence.

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.

Chapter 3

Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning

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.

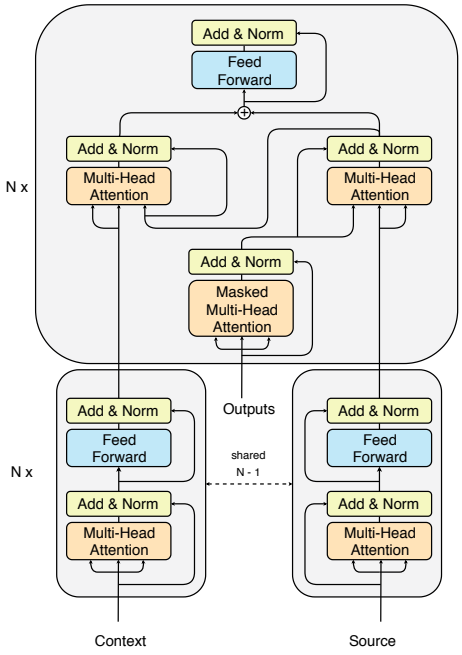


Figure 1: Context-aware model

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-

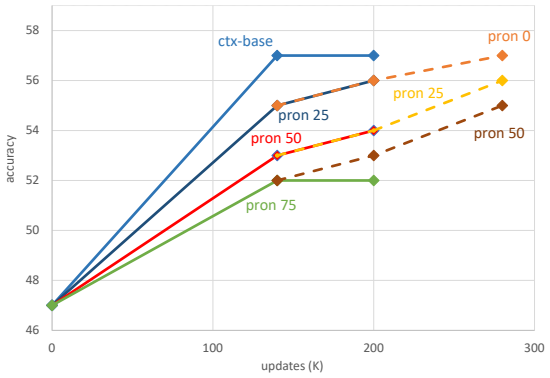


Figure 2: Challenge set accuracy. Full lines show fine-tuning from the baseline and dashed lines from a previous oracle model. Fine-tuning with a lr=10^{-4}.

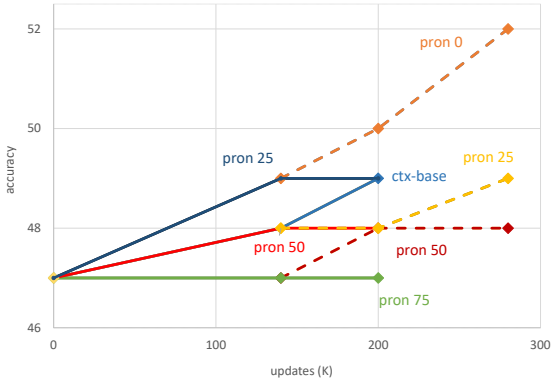


Figure 3: Challenge set accuracy. lr=10^{-5}.

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

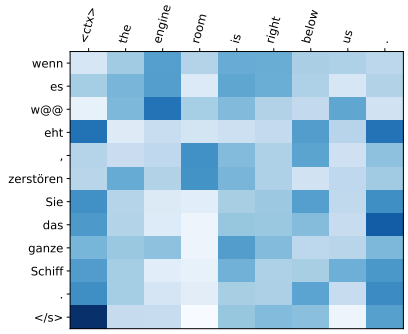


Figure 4: Context attention layer 1

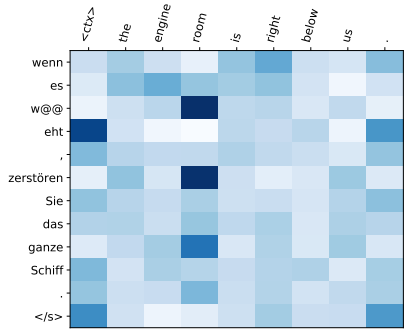


Figure 5: Context attention layer 2

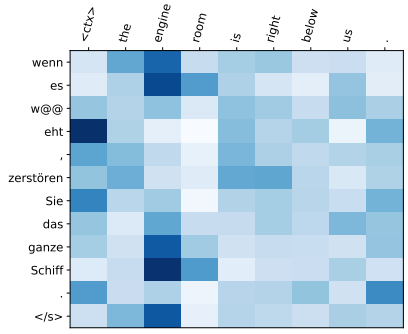


Figure 6: Context attention layer 3

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

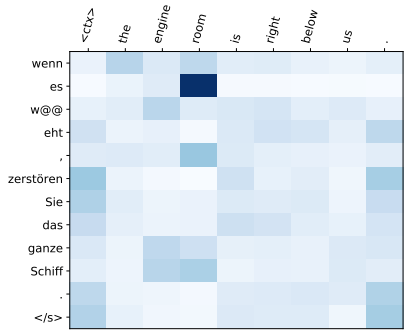


Figure 7: Context attention layer 6

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*.

Errata

The published version had a mistake in Figure 4. This version of the paper has the correct figure.

Chapter 4

Addressing Zero-Resource Domains Using Document-Level Context in Neural Machine Translation

Addressing Zero-Resource Domains Using Document-Level Context in Neural Machine Translation

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

Abstract

Achieving satisfying performance in machine translation on domains for which there is no training data is challenging. Traditional supervised domain adaptation is not suitable for addressing such zero-resource domains because it relies on in-domain parallel data. We show that when in-domain parallel data is not available, access to document-level context enables better capturing of domain generalities compared to only having access to a single sentence. Having access to more information provides a more reliable domain estimation. We present two document-level Transformer models which are capable of using large context sizes and we compare these models against strong Transformer baselines. We obtain improvements for the two zero-resource domains we study. We additionally provide an analysis where we vary the amount of context and look at the case where in-domain data is available.

1 Introduction

Training robust neural machine translation models for a wide variety of domains is an active field of work. NMT requires large bilingual resources which are not available for many domains and languages. When there is no data available for a given domain, e.g., in the case of web-based MT tools, this is a significant challenge. Despite the fact that these tools are usually trained on large scale datasets, they are often used to translate documents from a domain which was not seen during training. We call this scenario zero-resource domain adaptation and present an approach using document-level context to address it.

When an NMT model receives a test sentence from a zero-resource domain, it can be matched to similar domains in the training data. This is to some extent done implicitly by standard NMT. Alternatively, this matching can be facilitated by a domain adaptation technique such as using special domain tokens and features (Kobus et al., 2017; Tars and Fishel, 2018). However, it is not always easy to determine the domain of a sentence without larger context. Access to document-level context makes it more probable that domain signals can be observed, i.e., words representative of a domain are more likely to be encountered. We hypothesize that this facilitates better matching of unseen domains to domains seen during training and provide experimental evidence supporting this hypothesis.

Recent work has shown that contextual information improves MT (Miculicich et al., 2018; Voita et al., 2019b; Maruf et al., 2019), often by improving anaphoric pronoun translation quality, which can be addressed well with limited context. However, in order to address discourse phenomena such as coherence and cohesion, access to larger context is preferable. Voita et al. (2019b,a) were the first to show large improvements on lexical cohesion in a controlled setting using challenge sets. However, previous work did not make clear whether previous models can help with disambiguation of polysemous words where the sense is domain-dependent.

In this work, we study the usefulness of document-level context for zero-resource domain adaptation (which we think has not been studied in this way before). We propose two novel Transformer models which can efficiently handle large context and test their ability to model multiple domains at once. We show that document-level models trained on multi-domain datasets provide improvements on zero-resource domains. We evaluate on English→German translation using TED and PatTR (patent descriptions) as zero-resource domains. In addition to measuring translation quality, we conduct a manual evaluation targeted at word disambiguation. We also present additional experiments on classical domain adaptation where access to in-domain TED and PatTR data is allowed.

Our first proposed model, which we call the domain embedding model (DomEmb) applies average or max pooling over all context embeddings and adds this representation to each source token-level embedding in the Transformer. The second model is conceptually similar to previous work on context-aware NMT (Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Zhang et al., 2018) and introduces additional multi-head attention components in the encoder and decoder in order to handle the context. However, in order to facilitate larger context sizes, it creates a compressed context representation by applying average or max pooling with a fixed window and stride size. We compare our proposed models against previous context-aware NMT architectures and techniques for handling multi-domain setups, and show they improve upon strong baselines. The proposed models encode context in a coarse-grained way. They only have a limited ability to model discourse phenomena such as coreference resolution, so the gains we see in a multi-domain setup show that they encode domain information. Evaluating on multiple and zero-resource domains allows us to show that context can be used to capture domain information.

The contributions of our work can be summarized as follows: we (i) propose two NMT models which are able to handle large context sizes, (ii) show that document-level context in a multi-domain experimental setup is beneficial for handling zero-resource domains, (iii) show the effect of different context sizes and (iv) study traditional domain adaptation with access to in-domain data.

2 Related Work

Domain adaptation Several previous works address the problem that standard NMT may fail to adequately model all domains in a multi-domain setup even when all of the domains are known in advance. Kobus et al. (2017) introduce using domain tags for this problem, a similar method to the domain embedding model in our paper. These domain tags are mapped to corresponding embeddings and are either inserted at the beginning of the sentence or concatenated to the token-level embeddings. The domain embeddings are reserved for specific domains and are fixed for all sentences in a given domain. The number of distinct domain embeddings is limited to the number of known domains. Tars and Fishel (2018) define a similar approach which uses oracle domain tags and tags obtained using supervised methods and unsupervised clustering. However, clustering limits how many domains can be taken into consideration. Furthermore, this approach assumes that sufficient domain information can be obtained from a single sentence alone. Document-level classifiers (Xu et al., 2007) address this problem, but they are not jointly trained with the MT model. Further work in multi-domain MT is Foster and Kuhn (2007) who propose mixture models to dynamically adapt to the target domain, Foster et al. (2010) who build on this work and include instance weighting, Zeng et al. (2018) where domain-specific and domain-shared annotations from adversarial domain classifiers are used and Britz et al. (2017) where a discriminator is used to backpropagate domain signals.

Continued training is an established technique for domain adaptation if access to in-domain resources is possible. The method entails initially training on out-of-domain data, and then continuing training on in-domain data (Luong and Manning, 2015). Chen et al. (2017) and Zhang and Xiong (2018) improve upon this paradigm by integrating a domain classifier or a domain similarity metric into NMT and modifying the training cost based on weights indicating in-domain or out-of-domain data. Sajjad et al. (2017) and Farajian et al. (2017) use continued training in a multi-domain setup and propose various ways of fine-tuning to in-domain data. Standard continued training (Luong and Manning, 2015) leads to catastrophic forgetting, evident by the degrading performance on the out-of-domain dataset. Freitag and Al-Onaizan (2016) address this issue by ensembling the original and the fine-tuned model. We show that our model obtains significant improvements compared to a baseline with the ensembling paradigm. In contrast to these previous works, we do not know the domains during training. Our proposed approaches model the domain implicitly by looking at document-level context. Moreover, we evaluate performance on domains not seen during training.

Naradowsky et al. (2020) adapt to unseen domains using bandit learning techniques. The method relies on explicit user feedback which is not always easily available. Bapna and Firat (2019) propose a retrieval-based method that, at inference time, adapts to domains not seen during training. However, they assume access to in-domain parallel data at inference time, and they retrieve parallel phrases from this in-domain data. In our

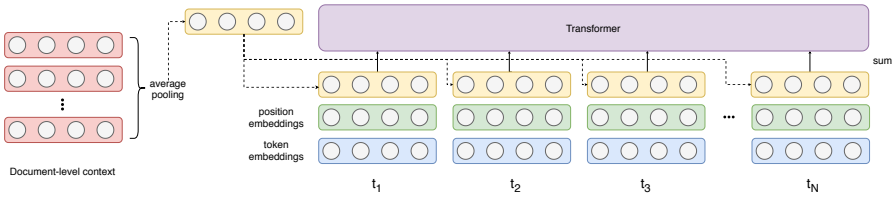


Figure 1: Domain embedding Transformer.

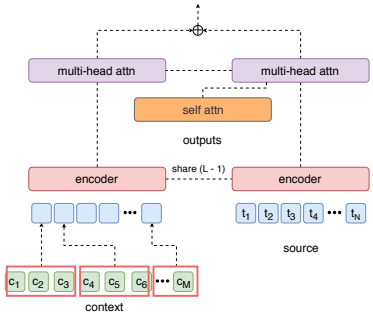


Figure 2: Context-aware Transformer with pooling.

zero-resource experiments, we have no access to in-domain parallel data.

Context-aware NMT A separate field of inquiry is context-aware NMT which proposes integrating cross-sentence context (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Tu et al., 2018; Maruf and Haffari, 2018; Voita et al., 2019b; Maruf et al., 2019; Yang et al., 2019; Voita et al., 2019a; Tan et al., 2019). These works show that context helps with discourse phenomena such as anaphoric pronouns, deixis and lexical cohesion. Kim et al. (2019) show that using context can improve topic-aware lexical choice, but in a single-domain setup.

Previous work on context-aware NMT has mostly worked with limited context. Miculicich et al. (2018) address the problem by reusing previously computed encoder representations, but report no BLEU improvements by using context larger than 3 sentences. Zhang et al. (2018) find 2 sentences of context to work the best. Maruf and Haffari (2018) use a fixed pretrained RNN encoder for context sentences and only train the document-level RNN. Junczys-Dowmunt (2019) concatenates sentences into very large inputs and outputs as in Tiedemann and Scherrer (2017). Maruf et al. (2019) propose a scalable context-aware model by using sparsemax which can ignore certain words and hierarchical attention which first computes sentence-level attention scores and subsequently word-level scores. However, for domain adaptation, the full encoder representation is too granular and not the most efficient way to obtain domain signals, for which we present evidence in our experiments. Stojanovski and Fraser (2019a); Macé and Servan (2019) propose a similar approach to our domain embedding model, but they do not investigate it from a domain adaptation perspective.

To our knowledge, our work is the first at the intersection of domain adaptation and context-aware NMT and shows that document-level context can be used to address zero-resource domains.

3 Model

The models we propose in this work are extensions of the Transformer (Vaswani et al., 2017). The first approach introduces separate domain embeddings applied to each token-level embedding. The second is conceptually based on previous context-aware models (Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Zhang et al., 2018). Both models are capable of handling document-level context. We modify the training data so that all sentences have access to the previous sentences within the corresponding source document. Access to the document-level context is available at test time as well. Sentences are separated with a special <SEP> token from the next sentence. We train and evaluate our models with a 10 sentence context.

3.1 Domain Embedding Transformer

The first model is shown in Figure 1. It is inspired by Kobus et al. (2017) which concatenates a special domain tag to each token-level embedding. Kobus et al. (2017) assume access to oracle domain tags during training. However, at inference, perfect domain knowledge is not possible. Consequently, the domain has to be predicted in advance which creates a mismatch between training and inference. An additional problem is inaccurately predicted do-

main tags at test time. We modify this approach by replacing the predefined special domain tag with one inferred from the document context. A disadvantage of this approach as opposed to Kobus et al. (2017) is that there is no clear domain indicator. However, the model is trained jointly with the component inferring the domain which increases the capacity of the model to match a sentence from an unseen domain to a domain seen during training.

The main challenge is producing the domain embedding from the context. We use maximum (DomEmb(max)) or average pooling (DomEmb(avg)) over all token-level context embeddings, both resulting in a single embedding representation. We do not apply self-attention over the context in this model. The intuition is that the embeddings will contain domain information in certain regions of the representation and that this can be extracted by max or average pooling. More domain-specific words will presumably increase the related domain signal. In contrast to a sentence-level model, large context can help to more robustly estimate the domain. Based on preliminary experimental results, we add a feed-forward neural network after the pooled embedding representation in DomEmb(avg), but not in DomEmb(max). We represent each token as a sum of positional, token-level embeddings and the inferred domain embedding. As the model only averages embeddings, the computational overhead is small. A computational efficiency analysis is provided in the appendix.

3.2 Context-Aware Transformer with Pooling

The second approach (CtxPool) is similar to previous work on context-aware NMT (e.g., (Stojanovski and Fraser, 2018; Zhang et al., 2018)). The model is outlined in Figure 2. It first creates a compact representation of the context by applying max or average pooling over the context with certain window and stride sizes. The intuition is similar to DomEmb, but pooling over a window provides a more granular representation. We use the concatenation of all context sentences (separated by <SEP>) as input to CtxPool.

The output of applying max or average pooling over time is used as a context representation which is input to a Transformer encoder. We share the first $L-1$ encoder layers between the main sentence and the context. L is the number of encoder layers. In the decoder, we add an additional multi-head attention (MHA) over the context. This attention is conditioned on the MHA representation from the main sentence encoder. Subsequently, these two representations are merged using a gated sum. The gate controls information flow from the context.

In contrast to DomEmb, CtxPool can be used to handle other discourse phenomena such as anaphora resolution. In this work, we use a window size of 10, suitable for domain adaptation. For anaphora, summarizing ten neighboring words makes it difficult to extract antecedent relationships. Careful tuning of these parameters in future work may allow modeling both local and global context.

4 Experiments

4.1 Experimental Setup

We train En→De models on Europarl, NewsCommentary, OpenSubtitles, Rapid and Ubuntu. TED and PatTR are considered to be zero-resource domains for which we have no parallel data. In additional experiments, we also consider classical domain adaptation where we do use TED and PatTR parallel data in a continued training setup. The models are implemented in Sockeye (Hieber et al., 2017). The code and the datasets are publicly available.[1] The preprocessing details and model hyperparameters are provided in the appendix.

4.2 Datasets

The datasets for some domains are very large. For example, OpenSubtitles contains 22M sentences and PatTR 12M. Due to limited computational resources, we randomly sample documents from these domains, ending up with approximately 10% of the initial dataset size. We keep the original size for the remaining datasets. Dataset sizes for all domains are presented in Table 1. The development and test sets are also randomly sampled from the original datasets. We sample entire documents rather than specific sentences. For TED we use tst2012 as dev and tst2013 as test set. The TED and PatTR dev sets are only used in the fine-tuning experiments where we assume access to in-domain data and are not used in any other experiment.

Europarl, NewsCommentary, OpenSubtitles, Rapid and TED are provided with document boundaries. Ubuntu lacks a clear discourse structure and PatTR is sentence-aligned, but provides document IDs. Previous work has shown that context-aware NMT performance is not significantly degraded

[1] https://www.cis.uni-muenchen.de/~dario/projects/zero_domain

domain	train	dev	test
Europarl	1.8M	3.2K	3.0K
NewsCommentary	0.3M	1.5K	1.5K
OpenSubtitles	2.2M	2.7K	3.3K
Rapid	1.5M	2.5K	2.5K
Ubuntu	11K	1.1K	0.6K
TED	0.2M	1.7K	1.0K
PatTR	1.2M	2.0K	2.2K

Table 1: Domain datasets sizes in sentences.

from lack of document boundaries (Müller et al., 2018; Stojanovski and Fraser, 2019b) or random context (Voita et al., 2018). To a large extent, both issues can be ignored, given the nature of our models. DomEmb is oblivious to the sentence order. CtxPool preserves some notion of sequentiality, but it should also be robust to these issues. Furthermore, we focus on obtaining domain signals. Even in an extreme case where the context comes from a different document (but from the same domain) we hypothesize similar performance. We later conduct an ablation study into whether arbitrary context from the same domain has a negative effect on performance. The results partially support our hypothesis by either matching or exceeding sentence-level performance, but also show that the correct context is important to obtain the best results.

4.3 Baselines

We compare our proposed methods against a sentence-level baseline (SentBase) and the domain tag (TagBase) approach (Kobus et al., 2017). We train TagBase with oracle domain tags, while at test time, we use tags obtained from a document-level domain classifier. All sentences within a document are marked with the same predicted domain tag. The domain classifier is a two-layer feed-forward network and the documents are represented as a bag-of-words. The classifier obtains an accuracy of 98.6%. By design, documents from TED and PatTR were marked with tags from the remaining domains. Additionally, we compare with a context-aware model (CtxBase) which is similar to CtxPool, but we feed the full context to the context Transformer encoder, without applying max or average pooling beforehand. This model has token-level granular access to the context. We also train a concatenation model (ConcBase) (Tiedemann and Scherrer, 2017) using source-side context.

5 Results

5.1 Zero-Resource Domain Adaptation

In zero-resource domain adaptation experiments, we do not use any data from TED or PatTR, neither as training nor development data. The models are trained on our multi-domain dataset consisting of five domains. The results are shown in Table 2. We compute statistical significance with paired bootstrap resampling (Koehn, 2004).

SentBase achieves 16.7 and 32.9 BLEU on PatTR and TED respectively. The domains seen during training are more similar to TED in comparison to PatTR which is the reason for the large BLEU score differences. Our proposed models improve on PatTR by up to 0.4 BLEU and on TED by up to 1.0 BLEU. Improvements vary, but all models increase the BLEU score. The TagBase model does not improve significantly over SentBase.

	PatTR	TED
SentBase	16.7	32.9
TagBase	16.8	33.0
DomEmb(max)	**17.1**†	**33.9**†
DomEmb(avg)	**17.1**†	33.8†
CtxPool(max)	16.9	33.6‡
CtxPool(avg)	**17.1**†	**33.9**†

Table 2: Results on zero-resource domain adaptation for PatTR and TED. Best results in bold. †- statistical significance with $p < 0.01$, ‡- $p < 0.05$.

Our document-level models are robust across the two domains. These results confirm our assumption that access to document-level context provides for a domain signal. These models are oblivious to the actual characteristics of the domain since it was not seen in training, but presumably, they managed to match the zero-resource domain to a similar one. We assume that the reason for the larger improvements on TED in comparison to PatTR is that TED is a more similar domain to the domains seen in training. As a result, matching TED to seen domains was easier for all models. Table 2 shows that our proposed models improve on PatTR and TED and provides evidence that document-level context is useful for addressing zero-resource domains.

5.2 Evaluating Domains Seen During Training

We assume that the improvements on zero-resource domains are because of document-level models having an increased capability to model domain.

domain	SentBase	TagBase	DomEmb(max)	DomEmb(avg)	CtxPool(max)	CtxPool(avg)
Europarl	31.3	31.4	32.3†	**32.5**†	32.4†	32.3†
NewsComm	32.8	32.6	32.7	33.0	**33.1**‡	32.8
OpenSub	26.6	27.1‡	27.0‡	**27.5**†	27.3†	27.4†
Rapid	40.7	40.9	41.1‡	41.5†	41.4†	**41.6**†
Ubuntu	31.5	**34.6**†	32.8‡	31.9	31.6	32.1
Average	30.4	30.9	**31.0**	**31.0**	30.9	**31.0**
Joint	29.1	29.2	29.5†	**29.8**†	29.7†	**29.8**†

Table 3: Results on the multi-domain dataset. Joint and average scores including PatTR and TED. Statistical significance computed for all scores except for Average. †- $p < 0.01$, ‡- $p < 0.05$.

As a result, we also evaluate on the other domains which were seen during training. We show average BLEU and the BLEU score on the concatenation of all test sets. This is a useful way to evaluate in a multi-domain setting because it is less sensitive to larger improvements on a smaller test set.

Table 3 shows the results. We first compare the baseline against DomEmb(avg). The smallest improvement is on NewsCommentary, only 0.2 BLEU. Improvements vary between 0.8 and 1.2 BLEU on Europarl, OpenSubtitles and Rapid. On Ubuntu, this model improves only by 0.4 BLEU. Joint and average BLEU improve by 0.7 and 0.6, respectively. Replacing average pooling with maximum pooling leads to slightly worse results on all domains except Ubuntu, but still improves upon the baseline. Our assumption is that averaging handles situations when there is a mix of domain signals because it can emphasize the more frequent domain signals. Max pooling is not able to differentiate between less and more frequent domain signals.

CtxPool(avg) and DomEmb(avg) perform similarly and have the same average and joint BLEU scores. Max pooling is slightly worse as shown by the performance of CtxPool(max). TagBase is not very effective in our experiments, improving slightly on some domains and only performing well on Ubuntu. We show that document-level context is useful for modeling multiple known domains at the same time. In the appendix we show translation examples from SentBase and DomEmb(avg).

5.3 Context Length

We also investigate the effect of context size on DomEmb(avg). Previous work on context-aware NMT (Zhang et al., 2018; Miculicich et al., 2018) typically showed that large context fails to provide for consistent gains. But this applies to more granular models which resemble the context-aware baseline CtxBase. In contrast, we observe that larger context does provide for improvements. We assume that for DomEmb, access to more context improves the likelihood of encountering domain-specific tokens.

domain	ctx=1	ctx=5	ctx=10
Europarl	31.5	32.0†‡	**32.5**†♣
NewsComm	32.7	32.9	**33.0**
OpenSub	26.8	27.2*‡	**27.5**†✧
Rapid	41.1‡	**41.5***‡	**41.5***
Ubuntu	32.5	**32.9***‡	31.9
PatTR	17.0‡	**17.2**‡	17.1
TED	33.5**	33.7‡	**33.8**
Average	30.7	**31.1**	31.0
Joint	29.3‡	29.7†‡	**29.8**†♣

Table 4: Results using the DomEmb(avg) model with different context sizes. Context size in number of previous sentences. ‡- $p < 0.01$, ** - $p < 0.05$, compared to SentBase. †- $p < 0.01$, * - $p < 0.05$, compared to *ctx=1*. ♣ - $p < 0.01$, ✧ - $p < 0.05$, compared to *ctx=5*.

We compare different context sizes and show the results in Table 4. A context size of 1 (*ctx=1*) obtains the lowest scores on all domains. Using *ctx=5* is comparable or slightly worse than *ctx=10*. Both *ctx=1* and *ctx=5* get higher scores on Ubuntu and obtain significant improvements over SentBase on the full test set. Significance indicators for *ctx=10* compared with respect to SentBase were already presented in Table 3. Due to resource limitations, we do not conduct a similar study for CtxPool.

5.4 Comparison to Context-Aware Baselines

Previous work on context-aware NMT has shown improvements in single-domain scenarios. In our work, we put two context-aware models to the test in a multi-domain setup. All models are trained with a 5 sentence context. The results in Table 5 show that all models improve to varying degrees. They perform similarly on NewsCommentary and

OpenSubtitles. CtxBase and ConcBase obtain better results on Europarl than DomEmb(avg) and worse on Ubuntu. CtxBase is best on Rapid. Both baselines obtained better scores on TED, showing they have some capacity to transfer to unseen domains. However, both failed to improve on PatTR.

domain	CtxBase	ConcBase	DomEmb(a)
Europarl	**32.4**†	**32.4**†	32.0†
NewsCo	32.8	32.7	**32.9**
OpenSub	27.2‡	**27.4**†	27.2†
Rapid	**41.8**†	40.8	41.5†
Ubuntu	31.6	29.1	**32.9**†
PatTR	16.6	14.8	**17.2**†
TED	**34.1**†	**34.1**†	33.7†
Average	30.9	30.2	**31.1**
Joint	**29.7**†	29.5	**29.7**†

Table 5: Comparison with the context-aware baseline CtxBase and the concatenation model ConcBase. †- $p < 0.01$, ‡- $p < 0.05$ compared to SentBase.

We use 5 sentences of context for this experiment. Scaling the baseline models to large context is challenging with regards to computational efficiency and memory usage. In contrast, DomEmb scales easily to larger context. Furthermore, our analysis shows that DomEmb(avg) has the best average and joint score (CtxBase obtains the same joint score), improves on both unseen domains and consistently obtains significant improvements on all domains except NewsCommentary. As previous works show (Müller et al., 2018), these context-aware baselines improve fine-grained discourse phenomena such as anaphora resolution. We show in our manual analysis that DomEmb(avg) does not improve anaphoric pronoun translation which indicates that the improvements of our proposed model and the context-aware baselines are orthogonal.

5.5 Translation of Domain-Specific Words

We also evaluated the translation of domain-specific words. We extracted the most important words from a domain based on TF-IDF scores and selected the top 100 with the highest scores which have more than 3 characters. Next, we follow Liu et al. (2018) and compute alignments using *fastalign* (Dyer et al., 2013) based on the training set and force align the test set source sentences to the references and generated translations. We then compute the F_1 score of the translation of the domain-specific words. Results are shown in Table 6. We compare SentBase with DomEmb(avg).

	SentBase	DomEmb(avg)
Europarl	0.661	**0.667**
NewsComm	0.649	**0.650**
OpenSub	0.435	**0.453**
Rapid	0.724	**0.730**
Ubuntu	0.434	**0.439**
PatTR	0.407	**0.409**
TED	0.551	**0.565**

Table 6: F_1 score for domain-specific words.

domain	SentBase	DomEmb(a)
PatTR	34.4	34.4
	ensemble	
Europarl	29.0	**29.6**†
NewsCommentary	28.7	**28.9**
OpenSubtitles	22.8	**23.4**†
Rapid	35.1	**35.7**†
Ubuntu	33.0	**33.4**
PatTR	29.2	**29.4**
TED	29.8	**30.4**‡
Average	29.7	**30.1**
Joint	30.2	**30.6**†

Table 7: Domain adaptation results on PatTR for SentBase and DomEmb(avg). †- $p < 0.01$, ‡- $p < 0.05$.

domain	SentBase	DomEmb(a)
TED	36.1	**36.6**‡
	ensemble	
Europarl	30.4	**30.8**†
NewsCommentary	31.9	**32.2**‡
OpenSubtitles	24.6	**25.4**†
Rapid	38.8	**39.5**†
Ubuntu	**32.7**	32.4
PatTR	16.9	**17.0**‡
TED	35.4	**35.8**‡
Average	30.1	**30.4**
Joint	28.4	**28.8**†

Table 8: Domain adaptation results on TED for SentBase and DomEmb(avg). †- $p < 0.01$, ‡- $p < 0.05$.

DomEmb(avg) improved the F_1 score across all domains with the largest improvements on OpenSubtitles and TED. Our assumption is that the baseline translation of OpenSubtitles domain-specific words is more formal. A large part of the seen domains contain formal language in contrast to the

domain	Europarl	NewsComm	OpenSub	Rapid	Ubuntu	PatTR	TED	True
Europarl	**31.3**	30.1	30.6	30.3	30.7	30.7	30.7	32.5
NewsComm	30.6	**32.8**	31.9	30.1	32.3	31.5	32.1	33.0
OpenSub	22.2	23.1	**27.1**	22.0	25.4	24.4	26.7	27.5
Rapid	39.5	37.0	38.7	**41.3**	40.3	40.4	38.9	41.5
Ubuntu	29.3	29.1	29.2	29.6	**31.4**	31.1	30.1	31.9
PatTR	16.6	16.2	16.3	16.5	16.9	**17.1**	16.8	17.1
TED	30.0	33.0	33.1	28.8	33.4	31.5	**33.7**	33.8

Table 9: Results from the ablation study investigating the influence of context from a different domain. Each row shows which domain is used as the test set and each column shows from which domain the context originates.

informal subtitles. Lack of context seems to have biased SentBase to generate more formal translations. We later conduct a manual analysis on the TED test set where we confirm that word sense disambiguation is indeed improved in DomEmb(avg).

5.6 Domain Adaptation with Available In-Domain Data

We also conduct a classical domain adaptation evaluation where access to in-domain data is allowed. We either use PatTR or TED as in-domain data and evaluate with SentBase and DomEmb(avg). In both cases we consider the concatenation of the remaining domains as out-of-domain. This setup differs from zero-resource domain adaptation because we assume access to in-domain training and dev data.

First, we train the baseline and DomEmb(avg) on out-of-domain data. Since these initial models are identical to the ones in the zero-resource setup, we reuse them. We then continue training on the corresponding in-domain data. Table 7 shows the results for PatTR. Fine-tuning the baseline and DomEmb(avg) on PatTR improves BLEU by a large margin, both obtaining 34.4 BLEU. The results are unsurprising because our model is tailored to multi-domain setups and is unlikely to contribute to large improvements when fine-tuning on a single domain. Identifying the domain in such a case is trivial and using large context should not be helpful.

The strengths of our approach come to light by comparing it against SentBase in an ensembling scenario as in Freitag and Al-Onaizan (2016). We ensemble DomEmb(avg) trained on out-of-domain data with DomEmb(avg) fine-tuned on in-domain data and do the same for SentBase. The DomEmb(avg) ensemble is better than the SentBase ensemble on all domains and on joint BLEU. Similar results are obtained when fine-tuning on TED which are shown in Table 8.

5.7 Ablation

We previously hypothesized that our models will benefit from context from different documents within the same domain. We conduct an ablation study to test this assumption using DomEmb(avg) model, similar to the study in (Kobus et al., 2017), where they investigated the effect of giving the wrong domain tag to every sentence.

For DomEmb(avg), we simulate this approach by replacing the real contextual representation of each test sentence with C_d, which is context representative of domain d. We first compute $C_d^{'} = \frac{1}{N_d} \sum_{i=1}^{N_d} c_i^d$ where c_i^d is the contextual representation of a test sentence in domain d and N_d is the number of test sentences in d. c_i^d is the average of the context token-level embeddings for sentence i. Finally, $C_d = \arg\max_{c_i^d} cos(c_i^d, C_d^{'})$. This procedure is conducted for each domain d separately.

Table 9 shows the results. On OpenSubtitles, Rapid, PatTR and TED, DomEmb(avg) improves on the sentence-level baseline if presented with context from the same domain (which is usually not from the same document). On Europarl, NewsCommentary and Ubuntu, it performs similarly to the baseline. In almost all cases, providing a mismatched context degrades the performance of the original DomEmb(avg). The results show that the model is relatively robust to incorrect but closely related context which provides evidence for our hypothesis that DomEmb captures domain-relevant features. However, the correct context is important to obtain the best results across all domains. Our finding is in contrast with recent results (Li et al., 2020) where they show that multi-encoder context-aware NMT models do not encode contextual information.

5.8 Manual Analysis

We conduct a manual analysis of SentBase and DomEmb(avg) by inspecting them on the TED test

set. We only consider translation differences related to word senses and ignore other types of mistakes. We find 156 cases where the two models translate a word in a different sense and at least one of them outputs the correct sense. We define 3 categories: (i) one model is correct while the other wrong; (ii) both are correct, but one is closer to the actual meaning and (iii) both are correct, but one matches the reference translation. DomEmb(avg) is better on (i) in 43 cases as opposed to the 19 cases where SentBase is better. The ratio of DomEmb(avg) being correct in contrast to SentBase is 23/12 in (ii) and 38/21 in (iii). This shows that DomEmb(avg) is better at coherence which is closely related to better domain modeling in multi-domain setups where the number of probable senses is larger than in a single domain. Furthermore, we find that DomEmb(avg) does not improve on pronoun translation. In fact, in several cases it introduced errors, thus ruling out better coreference resolution as a source of improvements.

6 Conclusion

We presented document-level context-aware NMT models and showed their effectiveness in addressing zero-resource domains. We compared against strong baselines and showed that document-level context can be leveraged to obtain domain signals. The proposed models benefit from large context and also obtain strong performance in multi-domain scenarios. Our experimental results show the proposed models obtain improvements of up to 1.0 BLEU in this difficult zero-resource domain setup. Furthermore, they show that document-level context should be further explored in future work on domain adaptation and suggest that larger context would be beneficial for other discourse phenomena such as coherence.

Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1).

References

Ankur Bapna and Orhan Firat. 2019. Non-Parametric Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective Domain Mixing for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126. Association for Computational Linguistics.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost Weighting for Neural Machine Translation Domain Adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-Domain Neural Machine Translation through Unsupervised Adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA. Association for Computational Linguistics.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. *CoRR*, abs/1612.06897.

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*.

Marcin Junczys-Dowmunt. 2019. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*

(Volume 2: Shared Task Papers, Day 1), pages 225–233, Florence, Italy. Association for Computational Linguistics.

Yunsu Kim, Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *DiscoMT@EMNLP*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378. INCOMA Ltd.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Frederick Liu, 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.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Valentin Macé and Christophe Servan. 2019. Using whole document context in neural machine translation. *arXiv preprint arXiv:1910.07481*.

Sameen Maruf and Gholamreza Haffari. 2018. Document Context Neural Machine Translation with Memory Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Jason Naradowsky, Xuan Zhang, and Kevin Duh. 2020. Machine translation system selection from bandit feedback. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 50–63, Virtual. Association for Machine Translation in the Americas.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural Machine Translation Training in a Multi-Domain Scenario. *CoRR*, abs/1708.08712.

Dario Stojanovski 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. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019a. Combining local and document-level context: The lmu munich neural machine translation system at wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 400–406, Florence, Italy. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019b. Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland. European Association for Machine Translation.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Sander Tars and Mark Fishel. 2018. Multi-Domain Neural Machine Translation. *arXiv preprint arXiv:1805.02282*.

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. 2018. Learning to Remember Translation History with a Continuous Cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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, Rico Sennrich, and Ivan Titov. 2019a. Context-Aware Monolingual Repair for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 876–885, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia.

Jia Xu, Yonggang Deng, Yuqing Gao, and Hermann Ney. 2007. Domain Dependent Statistical Machine Translation. In *MT Summit*.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing Context Modeling with a Query-Guided Capsule Network for Document-level Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 2018. Multi-Domain Neural Machine Translation with Word-Level Domain Context Discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. 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.

Shiqi Zhang and Deyi Xiong. 2018. Sentence Weighting for Neural Machine Translation Domain Adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190. Association for Computational Linguistics.

A Preprocessing and Hyperparameters

We tokenize all sentences using the script from Moses[2]. We apply BPE splitting[3] with 32K merge operations. We exclude TED and PatTR when computing the BPEs. The BPEs are computed jointly on the source and target data. Samples where the source or target are larger than 100 tokens are removed. We also apply a per-sentence limit of 100 tokens on the context, meaning that models trained on 10 sentences of context have a limit of 1000 tokens. A batch size of 4096 is used for all models.

We first train a sentence-level baseline until convergence based on early-stopping. All context-aware models are initialized with the parameters from this pretrained sentence-level baseline. Parameters that are specific to the models' architectures are randomly initialized. All proposed models in this work share the source, target, output and context embeddings. The models' architecture is a 6 layer encoder/decoder Transformer with 8 attention heads. The embedding and model size is 512 and the size of the feed-forward layers is 2048. The number of parameters for all models is shown in Table 10. We use label smoothing with 0.1 and dropout in the Transformer of 0.1. Models are trained on 2 GTX 1080 Ti GPUs with 11GB RAM.

Model	parameters
SentBase	61M
CtxBase	74M
CtxPool	74M
DomEmb(avg)	63M

Table 10: Number of model parameters. TagBase, ConcBase and DomEmb(max) have the same number of parameters as SentBase.

The initial learning rate for the document-level models is 10^{-4}. For the classical domain adaptation scenario with fine-tuning, we use a learning rate of 10^{-5} in order not to deviate too much from the well-initialized out-of-domain model. We lower the learning rate by a factor of 0.7 if no improvements are observed on the validation perplexity in 8 checkpoints. A checkpoint is saved every 4000 updates. We did not do any systematic hyperparameter search.

Before inference, we average the parameters of the 8 best checkpoints based on the validation perplexity. We use a beam size of 12. BLEU scores are computed on detokenized text using *multi-bleu-detok.perl* from the Moses scripts[4]. For the evaluation of translation of domain-specific words, we used the script from (Liu et al., 2018)[5].

B Datasets

We use the document-aligned versions of Europarl, NewsCommentary and Rapid from WMT 2019[6]. We also use OpenSubtitles[7][8] (Lison and Tiedemann, 2016), Ubuntu[9], PatTR[10] and TED[11].

C Validation performance

In Table 11, Table 12 and Table 13 we present BLEU scores on the development sets for all the experiments we ran. We only show results for the sets we actually used during training and therefore ignore TED and PatTR for which we had no access to data at training time. The results for TagBase are with oracle domain tags. For the experiments with continued training on TED and PatTR, we show results only on the development sets for TED and PatTR.

D Computational Efficiency

In this section, we compare the computational efficiency of our proposed methods. We compare how many seconds on average are needed to translate a sentence from the test set. The average times are 0.2588, 0.2763 ± 0.0124, 0.3662 for SentBase, DomEmb and CtxPool, respectively. DomEmb is insignificantly slower than the sentence-level baseline, in contrast to CtxPool, which is to be expected considering the additional applying of self-attention over the compressed context. In terms of training time, SentBase converged after 90 hours of training, DomEmb(avg) after 168h and CtxPool(avg) after 116h.

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[3] https://github.com/rsennrich/subword-nmt

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu-detok.perl

[5] https://github.com/frederick0329/Evaluate-Word-Level-Translation

[6] http://statmt.org/wmt19/translation-task.html

[7] http://opus.nlpl.eu/OpenSubtitles-v2018.php

[8] http://www.opensubtitles.org/

[9] http://opus.nlpl.eu/Ubuntu.php

[10] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

[11] https://wit3.fbk.eu/2015-01

domain	SentBase	TagBase	DomEmb(max)	DomEmb(avg)	CtxPool(max)	CtxPool(avg)
Europarl	33.3	33.6	33.6	33.7	33.8	33.8
NewsComm	34.1	34.3	34.1	34.1	34.2	34.1
OpenSub	33.3	34.2	34.2	34.5	34.1	34.2
Rapid	39.4	39.7	39.5	39.7	39.8	39.9
Ubuntu	40.2	43.0	41.3	42.6	42.0	42.2

Table 11: BLEU scores on the development sets of the multi-domain dataset.

domain	ctx=1	ctx=5	ctx=10
Europarl	33.5	33.8	33.7
NewsComm	34.0	34.2	34.1
OpenSub	33.7	34.1	34.5
Rapid	39.7	39.8	39.7
Ubuntu	41.5	43.0	42.6
domain	CtxBase	ConcBase	DomEmb(a)
Europarl	34.0	34.1	33.7
NewsComm	34.0	33.9	34.1
OpenSub	33.9	34.5	34.5
Rapid	40.1	39.1	39.7
Ubuntu	42.3	42.3	42.6

Table 12: Results on the development sets using the DomEmb(avg) model with different context sizes and comparing DomEmb(avg) with *ctx=10* against CtxBase and ConcBase.

domain	SentBase	DomEmb(a)
TED	33.2	33.4
PatTR	36.4	36.3

Table 13: Domain adaptation results on PatTR and TED for SentBase and DomEmb(avg) on the development sets.

E Examples

In Table 14 we show some example translations from the sentence-level baseline and our DomEmb(avg) model. We show examples where our model corrected erroneous translations from the baseline. Some of the proper translations should be evident from the main sentence itself, but some can only be inferred from context. The first four examples are from TED and the last from PatTR.

In the first example, we can see that the sentence-level baseline translates "students" as "Studenten" (university students), but the correct translation in this case is "Schüler" (elementary or high school student). The main sentence itself is not informative enough for the sentence-level model to make this distinction. In contrast, the DomEmb model has access to more information which provides for the appropriate bias towards the correct translation.

The second sentence depicts an example where it's nearly impossible for the baseline to make a correct prediction for the translation of "ambassador" because it depends on whether the person is male (Botschafter) or female (Botschafterin). In the third example, the sentence-level model translated "model" as in "a role model" (Vorbild), but the context indicates that the speaker talks about "fashion models".

Examples 4 and 5 are relatively unintuitive because the main sentences themselves should be enough to infer the correct translation. In example 4, "reflect" refers to the physical process of reflection and should not be translated as in "to reflect on oneself" ("denken"), while in example 5, "raise" refers to the action of "lifting" or "elevating"("aufwärtsbewegt" or "hochzuziehen") some object instead of "raising" as in "raising a plant (from a seed)" ("züchten").

The last example shows that the sentence-level model translates "springs" ("Federn" which is a part of the compound word "Druckfedern" in the reference) as in "water springs" ("Quellen" which is a part of the compound word "Kompressionsquellen") while it should be translated instead as in the physical elastic device. However, in other test sentences, both SentBase and DomEmb(avg) translated "spring" as a season, even though this should be less likely in PatTR, showing that our model does not always succeed in capturing domain perfectly.

Source We all knew we were risking our lives – the teacher, the **students** and our parents. *Reference* Wir alle wussten, dass wir unser Leben riskierten: Lehrer, **Schüler** und unsere Eltern. *SentBase* Wir alle wussten, dass wir unser Leben riskieren... den Lehrer, die **Studenten** und unsere Eltern. *DomEmb(avg)* Wir wussten alle, dass wir unser Leben riskierten. Der Lehrer, die **Schüler** und unsere Eltern.
Source That's why I am a global **ambassador** for 10x10, a global campaign to educate women. *Reference* Deshalb bin ich globale **Botschafterin** für 10x10, einer weltweiten Kampagne für die Bildung von Frauen. *SentBase* Aus diesem Grund bin ich ein globaler **Botschafter** für 10x10, eine weltweite Kampagne zur Ausbildung von Frauen. *DomEmb(avg)* Deshalb bin ich eine globale **Botschafterin** für 10x10, eine weltweite Kampagne zur Ausbildung von Frauen.
Source And I am on this stage because I am a **model**. *Reference* Und ich stehe auf dieser Bühne, weil ich ein **Model** bin. *SentBase* Und ich bin auf dieser Bühne, weil ich ein **Vorbild** bin. *DomEmb(avg)* Und ich bin auf dieser Bühne, weil ich ein **Model** bin.
Source It's going to bounce, go inside the room, some of that is going to **reflect** back on the door ... *Reference* Es wird abprallen, in den Raum gehen, ein Teil davon wird wieder zurück auf die Tür **reflektiert** ... *SentBase* Es wird abprallen, ins Zimmer gehen, etwas davon wird wieder an die Tür **denken** ... *DomEmb(avg)* Es wird abprallen, ins Zimmer gehen, etwas davon wird wieder über die Tür **reflektieren** ...
Source Tie member 60 is driven to **raise** movable cone 58 ... *Reference* Mit dem Zugelement 60 wird durch den An der bewegliche Kegel 58 **aufwärtsbewegt** ... *SentBase* Tie-Mitglied 60 wird angetrieben, bewegliche Konfitüre 58 zu **züchten** ... *DomEmb(avg)* Teemitglied 60 wird angetrieben, bewegliche Kegel 58 **hochzuziehen** ...
Source It is only when a certain pressure level is reached that the pistons are pushed back against the action of the **compression springs** ... *Reference* Erst bei Erreichen eines bestimmten Druckniveaus werden die Kolben gegen die Wirkung der **Druckfedern** zurückgeschoben ... *SentBase* Erst wenn ein gewisses Druckniveau erreicht ist, werden die Pistonen gegen die Wirkung der **Kompressionsquellen** zurückgedrängt ... *DomEmb(avg)* Erst wenn ein bestimmtes Druckniveau erreicht ist, werden die Pistonen gegen die Wirkung der **Kompressionsfedern** zurückgedrängt ...

Table 14: Example translations obtained using sentence-level baseline and the DomEmb(avg) model. Relevant parts of the examples are in bold.

Chapter 5

Combining Local and Document-Level Context: The LMU Munich Neural Machine Translation System at WMT19

Combining Local and Document-Level Context: The LMU Munich Neural Machine Translation System at WMT19

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

Abstract

We describe LMU Munich's machine translation system for English→German translation which was used to participate in the WMT19 shared task on supervised news translation. We specifically participated in the document-level MT track. The system used as a primary submission is a context-aware Transformer capable of both rich modeling of limited contextual information and integration of large-scale document-level context with a less rich representation. We train this model by fine-tuning a big Transformer baseline. Our experimental results show that document-level context provides for large improvements in translation quality, and adding a rich representation of the previous sentence provides a small additional gain.

1 Introduction

In this paper we describe the system we developed at the LMU Munich Center for Information and Language Processing, which we used to participate in the news translation task at WMT19. We submitted system runs for the English→German translation direction and specifically focus on the document-level translation track. The goal of the document-level track is to train machine translation models capable of taking into account larger context or even entire documents when translating sentences.

Supervised NMT has achieved state-of-the-art results (Bahdanau et al., 2015; Vaswani et al., 2017). Several works have claimed translation quality on a level similar to human translation. Wu et al. (2016) report translation quality on par with average bilingual human translators and Hassan et al. (2018) argue for parity to professional human translators on news translation from Chinese to English. However, these claims have been challenged in several ways with recent work (Läubli et al., 2018; Toral et al., 2018). One challenge is that these evaluations were done without giving evaluators access to the whole document-level context. They further show that human translations are preferred over automatic ones if evaluators are given document-level context. This is precisely the motivation for the document-level MT track in this year's WMT19.

One of the reasons for the failure of NMT in these context-dependent cases is not being able to model discourse-level phenomena. The straightforward reason for this is that traditional NMT does not have access to the context. As a result, it fails to account for several discourse-level phenomena, prominent ones being coreference resolution and coherence.

Coreference resolution has a particular impact on English→German translation, specifically for pronoun translation. English has only one third person singular pronoun that is routinely used for non-human references ("it"), while German has three, each representing a specific gender: masculine, feminine and neuter. Consider the following sentence: *We know it won't change students' behaviour instantly.* The translation of *it* into German can be, *er*, *sie* or *es* depending on the gender of the noun the English *it* is referencing. Since traditional NMT is working on the sentence-level, it has no way of ascertaining the appropriate gender and usually falls back to the data-driven prior, which is the neuter *es*.

Coherence is important in order to provide coherent translations across the whole given document. It is usually undesirable to produce translations with different meanings within a single document for the same ambiguous word.

Taking into account the whole document when generating translations will address some of the relevant discourse-level phenomena. An implicit effect that one could expect by modeling the whole

document is also modeling the underlying domain. On an abstract level, one can presume that this is happening in sentence-level models as well, however access to larger context is likely to improve the ability to implicitly identify the domain. Domain adaptation and multi-domain NMT have been extensively studied (Kobus et al., 2017; Freitag and Al-Onaizan, 2016; Farajian et al., 2017; Sajjad et al., 2017; Zhang and Xiong, 2018; Chen et al., 2017; Tars and Fishel, 2018). However, most previous works assume that the domain of each sentence is known at training time, which is often not the case.

Taking into consideration different discourse-level phenomena, we develop a Transformer (Vaswani et al., 2017) which can richly model the previous sentence, but also takes advantage of larger context. We borrow on previous work on context-aware NMT (Stojanovski and Fraser, 2018; Voita et al., 2018; Miculicich et al., 2018; Zhang et al., 2018) and add additional parameters in the encoder and decoder to account for the previous sentence. We limit the context since we want this part of the model to be able to do coreference resolution which very often can be addressed by looking at the first previous sentence. We additionally take the 10 previous sentences and create a simple document representation by averaging their embeddings. This embedding is subsequently added to each source token in the sentence to be translated in the same fashion as positional embeddings are added to the token-level embeddings in the Transformer. We assume that this representation can help provide a clear domain signal.

The remainder of the paper outlines the model in detail, and presents the experimental setup and obtained results.

2 Related Work

There are large number of works in NMT focusing on integrating document-level information into otherwise sentence-level models (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Tu et al., 2018; Maruf and Haffari, 2018). These works have shown that improvements in pronoun translation are achieved by better handling coreference resolution. Smaller improvements are observed for coherence and cohesion. The main intuition behind the models in these works is that they employ an additional encoder for contextual sentences and integrate the information in the encoder or decoder using a gating mechanism. Our model is similar to the context-aware Transformer models proposed in these works with some specifics which we discuss in Section 3.

We also extend the Transformer model with a simple document representation which we assume provides for a domain signal. This could be useful for domain disambiguation and improved coherence and cohesion. This model is similar to previous work on domain adaptation for NMT (Kobus et al., 2017; Tars and Fishel, 2018) where special domain tokens are either added to the beginning of the sentence or concatenated as additional features to the token-level embeddings. However, they assume a set of known domains in advance which is not the case in our work. We model the domain implicitly.

3 Model

In this work we develop two models: a previous-sentence and document-level context-aware Transformer. For our primary submission, we use a joint model combining both approaches into a single model. We use source side context only, both at training and testing time.

3.1 Previous-sentence context-aware Transformer

This context-aware model is in line with previous works on context-aware NMT (Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Zhang et al., 2018). The standard Transformer is extended to be able to receive an additional sentence as input. In this work we only use the first previous sentence. We feed this context sentence through the Transformer encoder. As suggested in Voita et al. (2018), we share the encoder for the main and context sentence. In order to provide information as to what is being encoded, we add a special token at the beginning of the context sentence. We share the encoder layers up to and including the penultimate layer. Unlike Voita et al. (2018), we do not integrate the context in the encoder, but rather in the decoder. As a result, the last encoder layer is the standard Transformer encoder, but it is not shared across the main and context sentence.

We modify the decoder by adding an additional

multi-head attention (MHA) sublayer on the context representation. As in the standard Transformer decoder layer, at training time, we first compute self-attention over the target sentence and use this to compute the MHA representation c_i over the main sentence. The output of this step is used to condition the MHA c_i^c over the context. Subsequently, the outputs of the MHA over the main and context representations, c_i and c_i^c, are merged using a gated sum. The use of the gate is similar to previous work (Wang et al., 2017; Voita et al., 2018). It is conditioned on c_i and c_i^c. The output is computed as follows:

$$s_i = g_i \otimes c_i + (1 - g_i) \otimes c_i^c$$

and the gate is computed as:

$$g_i = \sigma(W_e c_i + W_c c_i^c)$$

where σ represents sigmoid activation and $\otimes$ element-wise multiplication. The gate enables the model to control how much information should be used from the main sentence and from the context sentence. Finally, the output of the gated sum is passed through a feed-forward neural network.

3.2 Document-level context-aware Transformer

We also extend the model with the ability to consume larger context. Miculicich et al. (2018) proposed a model capable of using large context using hierarchical attention. They tackle the memory requirements of such models by reusing already computed sentence representations. This introduces limitations as to how the random batching usually used to train NMT works, since it is necessary to have the previous sentences of a given sentence in a document already processed. Furthermore, Miculicich et al. (2018) report that they fail to obtain significant improvements as the context increases. They do not improve results beyond context sizes of 2 or 3 sentences.

As a result, we make a simple modification to the Transformer which enables it to handle large context sizes. In this work we use up to 10 sentences of context, all of which are previous sentences (but it would also be possible to use the following sentences as well). We take the embeddings of all tokens within the context and simply average them. This averaged document representation is then passed through a feed-forward network. The final document-level representation is then added to all token-level source embeddings in the sentence to be translated in the same manner as the positional embeddings are added in the Transformer. A similar approach was proposed by Kobus et al. (2017) for domain adaptation in RNN-based NMT. The work differs since they have special tokens which indicate the domain and they concatenate them instead of adding them to the token-level embeddings. Our approach is more flexible since it only relies on having access to contextual information and does not require explicit domain knowledge. Our intuition with this approach is that the document representation should be informative of the type or domain of the document being translated.

We share all source, target, output and context embeddings. We freeze them in the continued training phase with the context-aware model in order for the model to be more memory efficient.

4 Experimental Setup

4.1 Preprocessing

The data is preprocessed by normalizing punctuation, tokenizing and truecasing with the scripts from Moses. We apply BPE splitting (Sennrich et al., 2016b) with 32K merge operations. BPE is computed jointly on both languages.

Corpus	sentences
CommonCrawl	2.1M x2
Europarl	1.5M x2
NewsCommentary	0.3M x2
Rapid	1.4M x2
WikiTitles	1.3M x2
ParaCrawl	13.5M
NewsCrawl	9.3M
NewsCrawl v2	16.9M

Table 1: Training data sizes after filtering. x2 - oversampling factor.

4.2 Data filtering

Samples where the length of the source, target or first previous sentence before BPE-splitting is over 50 tokens are removed. For the purposes of our document-level model, we also use larger context. In our experiments, we restrict the model to access only the 10 previous sentences at most. Samples where the total length of these sentences exceeds 500 are also removed. After applying BPE splitting, an additional length filtering step

is applied with a maximum length allowed of 100 for the source, target and first previous sentence. Document-level context is limited to 800.

WMT provides the large ParaCrawl corpus which is very noisy. In previous years at WMT, high scoring systems showed that it is necessary to perform aggressive filtering. We reuse some of the data selection steps proposed in Stahlberg et al. (2018). We run language identification and remove non-English and non-German sentences. Furthermore, all sentences are removed where one of the following conditions is met: a word is over 40 characters long, HTML tags in text, sentence length less than 4 words, character ratio between source and target sentence is over 1:3 or 3:1, source or target sentence is not identical after removing non-numerical characters and sentence does not end in a punctuation mark. As a result, the size of the ParaCrawl corpus was reduced from 30M to 13.5M sentences. Unfortunately, due to time constraints, we were not able to reproduce the data filtering and data selection suggested by Junczys-Dowmunt (2018) which obtained the top BLEU scores at WMT18. They showed that the optimal number of sentences is 8M. We assume that the higher number of presumably noisy sentences is affecting our initial baseline.

4.3 Backtranslation

As shown in previous years, using backtranslations (Sennrich et al., 2016a) is essential for strong translation quality. We train a German→English small Transformer and use it to backtranslate NewsCrawl data. Due to time constraints, we were not able to use the backtranslated data in the initial training of the English→German model. As a result, we fine-tune the already trained baseline with the backtranslated data mixed in with the parallel WMT data.

4.4 Hyperparameters

We train a big Transformer as a baseline. Embedding and hidden dimension size in the encoder and decoder is 1024. All attention sublayers use dot product attention and have 16 attention heads. The size of the feed-forward neural networks is 4096. The hidden dimension size of the context-aware encoder and context attention sublayer in the decoder is 512. All context-related attention sublayers have 8 attention heads. All models have 6 encoder and decoder layers. We use sinusoidal positional embeddings which are added to the token-level embeddings. In the case of the document-level model, we further add the average of all large-context embeddings. We apply residual dropout of 0.1 as in (Vaswani et al., 2017). Additionally, dropout of 0.1 is applied to the multi-head attention and feed-forward network. We also use label smoothing of value 0.1.

4.5 Training

We train the Transformer baseline with a warmup period and a learning rate of 10^{-4}. In all cases of continued training in the paper, we set the learning rate to 10^{-5}. We train the models with early-stopping based on the perplexity on the development set. We checkpoint the model every 4000 updates. The learning rate is reduced by a factor of 0.7 if no improvements are observed for 8 checkpoints. Training converges if no improvements are observed after 32 checkpoints. We train our context-aware models by continued training on the converged baseline. All parameters relating only to the context-aware parts of the architecture are randomly initialized. The batch size is set to 4096 tokens.

Model	parameters
baseline	217M
previous-sentence context	253M
document-level context	225M
joint model	261M

Table 2: Number of model parameters. All models are big Transformer models.

The number of parameters for all models are presented in Table 2. We train the models on 4 GTX 1080 Ti GPUs with 12GB RAM. We use Sockeye[1] (Hieber et al., 2018) to train the baseline and our context-aware models.

5 Empirical Evaluation

We present the results we obtain with our models in Table 3. We report results on the English→German newstest2017, newstest2018 and newstest2019. We report BLEU scores using sacreBLEU[2] (Post, 2018) on detokenized text. For the final submission, we processed quotation marks to match the German style.

We train our baseline on the data presented in Table 1. We initially train on the ParaCrawl

[1] https://github.com/awslabs/sockeye
[2] https://github.com/mjpost/sacreBLEU

dataset and an oversampled version of the other datasets. We train this baseline until convergence with early-stopping based on the perplexity on the development set. As a development set, we use newstest2018. After convergence, we fine-tune with 9.3M NewsCrawl backtranslations in addition to the dataset we used for the initial baseline. This baseline is used to initialize all the other context-aware models. It is interesting to observe that fine-tuning with NewsCrawl backtranslations and WMT data improves on newstest2017 and newstest2018, but significantly decreases the BLEU score on newstest2019.

	en→de		
Model	nt17	nt18	nt19
baseline	29.8	45.3	39.5
baseline*	30.3	45.6	38.5
previous-sentence*	30.5	46.0	38.6
document-level*	30.5	45.7	39.3
document-level	**31.1**	47.0	40.0
joint	**31.1**	**47.1**	**40.3**

Table 3: BLEU scores on newstest2017, newstest2018 and newstest2019. * - model trained with NewsCrawl backtranslations. All context-aware models fine-tuned on baseline*.

For training the context-aware models, we ignore the ParaCrawl data and use the remaining datasets. Depending on the setup, we either use the 16.9M NewsCrawl backtranslations with document boundaries or completely ignore them. Our previous sentence context-aware Transformer trained with NewsCrawl backtranslations do not provide for significant improvements. It increases the BLEU score from 38.5 to 38.6. However, the document-level model with averaging context embeddings obtains a BLEU score of 39.3.

We also remove the NewsCrawl backtranslations when fine-tuning our average context embedding Transformer. This proves to be very helpful and we manage to obtain 40.0 BLEU. It is interesting that this model also substantially improves the BLEU score on newstest2017 and newstest2018. One possible explanation of the adverse effect of using backtranslations is that our document-level model is more sensitive to noisy input. We leave a further examination of the issue for future work.

Finally, we train a joint model where we combine the average context embedding approach with the previous-sentence context-aware Transformer where we employ a separate encoder and modify the decoder. This further pushes the BLEU score to 40.3 on newstest2019 and slightly improves results on the other test sets. This is the system we used for the primary submission.

We also tried ensembling context-aware joint models. However, due to time constraints we only managed to train a single baseline. Therefore, all context-aware models were trained by fine-tuning on top of the single baseline. As a result, these models were not diverse enough and ensembling did not help. After the evaluation period, we also tried averaging the last 5 checkpoints of a single run of the joint model. This improved the score on newstest2019 to 40.8 BLEU.

6 Conclusion

In this work, we presented our system which we used to participate in the English→German news translation task at WMT19. We proposed two modifications to the standard Transformer architecture. We propose a context-aware Transformer which has a separate encoder and a modified decoder in order to provide for a fine-grained access to a limited context. We further extend this model by proposing to average the context token-level embeddings and add them to the main sentence embeddings. This enables access to large scale context. We show that the latter modification provides for large improvements with regards to a baseline and that combining both approaches leads to a further performance increase.

Acknowledgments

We would like to thank 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

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.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint arXiv:1803.05567*.

Felix Hieber, 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. Association for Machine Translation in the Americas.

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

Marcin Junczys-Dowmunt. 2018. Microsoft's submission to the WMT2018 news translation task: How i learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 429–434, Brussels, Belgium. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378. INCOMA Ltd.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

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

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

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. *CoRR*, abs/1708.08712.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Felix Stahlberg, Adri de Gispert, and Bill Byrne. 2018. The University of Cambridges machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 508–516, Brussels, Belgium. Association for Computational Linguistics.

Dario Stojanovski 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. Association for Computational Linguistics.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

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.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Zhaopeng Tu, 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.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. 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.

Shiqi Zhang and Deyi Xiong. 2018. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190. Association for Computational Linguistics.

Chapter 6

ContraCAT: Contrastive Coreference Analytical Templates for Machine Translation

ContraCAT: Contrastive Coreference Analytical Templates for Machine Translation

Dario Stojanovski[†*], Benno Krojer[†*], Denis Peskov[‡*], Alexander Fraser[†]
[†]Center for Information and Language Processing, LMU Munich
[‡]Computer Science, University of Maryland
{stojanovski,fraser}@cis.lmu.de, benno.krojer@gmail.com,
dpeskov@cs.umd.edu

Abstract

Recent high scores on pronoun translation using context-aware neural machine translation have suggested that current approaches work well. ContraPro is a notable example of a contrastive challenge set for English→German pronoun translation. The high scores achieved by transformer models may suggest that they are able to effectively model the complicated set of inferences required to carry out pronoun translation. This entails the ability to determine which entities could be referred to, identify which entity a source-language pronoun refers to (if any), and access the target-language grammatical gender for that entity. We first show through a series of targeted adversarial attacks that in fact current approaches are not able to model all of this information well. Inserting small amounts of distracting information is enough to strongly reduce scores, which should not be the case. We then create a new template test set ContraCAT, designed to individually assess the ability to handle the specific steps necessary for successful pronoun translation. Our analyses show that current approaches to context-aware NMT rely on a set of surface heuristics, which break down when translations require real reasoning. We also propose an approach for augmenting the training data, with some improvements.

1 Introduction

Machine translation is a complex task which requires diverse linguistic knowledge. The seemingly straightforward translation of the English pronoun *it* into German requires knowledge at the syntactic, discourse and world knowledge levels for proper pronoun coreference resolution (CR). The German third person pronoun can have three genders, determined by its antecedent: masculine (*er*), feminine (*sie*) and neuter (*es*). Previous work (Hardmeier and Federico, 2010; Miculicich Werlen and Popescu-Belis, 2017; Müller et al., 2018) proposed evaluation methods for pronoun translation. This has been of special interest in context-aware NMT models that are capable of using discourse-level information. Despite promising results (Bawden et al., 2018; Müller et al., 2018; Lopes et al., 2020), the question remains: Are transformers (Vaswani et al., 2017) truly *learning* this task, or are they exploiting simple heuristics to make a coreference prediction?

To empirically answer this question, we extend ContraPro (Müller et al., 2018)—a contrastive challenge set for automatic English→German pronoun translation evaluation—by making small adversarial changes in the contextual sentences. Our adversarial attacks on ContraPro show that context-aware transformer NMT models can easily be misled by simple and unimportant changes to the input. However, interpreting the results obtained from adversarial attacks can be difficult. The results indicate that NMT uses brittle heuristics to solve CR, but it is not clear what those heuristics are. In general, it is challenging to design attacks based on modifying ContraPro that can test specific phenomena that may be of interest.

[*]Equal contribution.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4732–4749
Barcelona, Spain (Online), December 8-13, 2020

Start:	The cat and the actor were hungry.
Original sentence	It (?) was hungrier.
Step 1:	The **cat** and the **actor** were hungry.
Markable Detection	It (?) was hungrier.
Step 2:	The cat and the actor were hungry.
Coreference Resolution	**It** (😺) was hungrier.
Step 3:	Der Schauspieler und die Katze waren hungrig.
Language Translation	Er / **Sie** (😺) / Es war hungriger.

Table 1: A hypothetical CR pipeline that sequentially resolves and translates a pronoun.

For this reason, we propose an *independent* set of templates for coreferential pronoun translation evaluation to systematically investigate which heuristics are being used. Inspired by previous work on CR (Raghunathan et al., 2010; Lee et al., 2011), we create a number of templates tailored to evaluating the specific steps of an idealized CR pipeline. We call this collection ContraCAT (😺), **Contra**stive **C**oreference **A**nalytical **T**emplates. The templates are constructed in a completely controlled manner, enabling us to easily create large number of coherent test examples and provide strong conclusions about the CR capabilities of NMT. The procedure we used in creating the templates can be adapted to many language pairs with little effort. Our 😺 results suggest that transformer models do not learn each step of a hypothetical CR pipeline.

We also present a simple data augmentation approach specifically tailored to pronoun translation. The experimental results show that this approach improves scores and robustness on some of our metrics, but it does not fundamentally change the way CR is being handled by NMT.

We publicly release ContraCAT and the adversarial modifications to ContraPro[1].

2 Coreference Resolution in Machine Translation

Addressing discourse phenomena is important for high-quality MT. Apart from document-level coherence and cohesion, anaphoric pronoun translation has proven to be an important testing ground for the ability of context-aware NMT to model discourse. Anaphoric pronoun translation is the focus of several works in context-aware NMT (Bawden et al., 2018; Voita et al., 2018; Stojanovski and Fraser, 2019; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

However, the choice of an evaluation metric for CR is nontrivial. BLEU-based evaluation is insufficient for measuring improvement in CR (Hardmeier, 2012) without carefully selecting or modifying test sentences for pronoun translation (Voita et al., 2018; Stojanovski and Fraser, 2018). Alternatives to BLEU include F_1, partial credit, and oracle-guided approaches (Hardmeier and Federico, 2010; Guillou and Hardmeier, 2016; Miculicich Werlen and Popescu-Belis, 2017). However, Guillou and Hardmeier (2018) show that these metrics can miss important cases and propose semi-automatic evaluation. In contrast, our evaluation is *completely* automatic.

We focus on scoring-based evaluation (Sennrich, 2017), which works by creating contrasting pairs and comparing model scores. Accuracy is calculated as how often the model chooses the correct translation from a pool of alternative incorrect translations. Bawden et al. (2018) manually create such a contrastive challenge set for English→French pronoun translation. ContraPro (Müller et al., 2018) follows this work, but creates the challenge set in an automatic way.

We show that making small variations in ContraPro substantially changes the scores. Our work is related to adversarial datasets for testing robustness used in Natural Language Processing tasks such as studying gender bias (Zhao et al., 2018; Rudinger et al., 2018; Stanovsky et al., 2019), natural language inference (Glockner et al., 2018) and classification (Wang et al., 2019).

Jwalapuram et al. (2019) propose a model for pronoun translation evaluation trained on pairs of sentences consisting of the reference and a system output with differing pronouns. However, as Guillou and Hardmeier (2018) point out, this fails to take into account that often there is not

[1] http://cistern.cis.lmu.de/contracat

a 1:1 correspondence between pronouns in different languages. As a result, a system translation may be correct despite not containing the exact pronoun in the reference, and incorrect even if containing the pronoun in the reference, because of differences in the translation of the referent. Moreover, introducing a separate model which needs to be trained before evaluation adds an extra layer of complexity in the evaluation setup and makes interpretability more difficult. In contrast, templates can easily be used to pinpoint specific issues of an NMT model. Our templates follow previous work (Ribeiro et al., 2018; McCoy et al., 2019; Ribeiro et al., 2020) where similar tests are proposed for diagnosing NLP models.

3 Do Androids Dream of Coreference Translation Pipelines?

Imagine a hypothetical coreference pipeline that generates a pronoun in a target language, as illustrated in Table 1. **First**, markables (entities that can be referred to by pronouns) are tagged in the source sentence (we restrict ourselves to concrete entities as we wish to detect gender). Then, the subset of animate entities are detected, and human entities are separated from other animate ones (since *it* cannot refer to a human entity). **Second**, coreferences are resolved in the source language. This entails handling phenomena such as world knowledge, pleonastic *it*, and event references. **Third**, the pronoun is translated into the target language. This requires selecting the correct gender given the referent (if there is one), and selecting the correct grammatical case for the target context (e.g., accusative, if the pronoun is the grammatical object in the target language sentence).

This idealized pipeline would produce the correct pronoun in the target language. The coreference steps resemble the rule-based approach implemented in Stanford CoreNLP's Coref-Annotator (Raghunathan et al., 2010; Lee et al., 2011). However, NMT models are currently unable to decouple the individual steps of this pipeline. We propose to isolate each of these steps through targeted examples.

4 Model

We use a transformer model for all experiments and train a sentence-level model as a baseline. The context-aware model in our experimental setup is a concatenation model (Tiedemann and Scherrer, 2017) (CONCAT) which is trained on a concatenation of consecutive sentences. CONCAT is a standard transformer model and it differs from the sentence-level model only in the way that the training data is supplied to it. The training examples for this model are modified by prepending the previous source and target sentence to the main source and target sentence, respectively. The previous sentence is separated from the main sentence with a special token <SEP>, on both the source and target side. This also applies to how we prepare the ContraPro and ContraCAT data. We train the concatenation model on OpenSubtitles2018 data prepared in this way. We remove documents overlapping with ContraPro. Preprocessing details and model hyper-parameters are presented in the Appendix.

5 Adversarial Attacks

5.1 About ContraPro

ContraPro is a contrastive challenge set for English→German pronoun translation evaluation. The set consists of English sentences containing an anaphoric pronoun "it" and the corresponding German translations. It contains three contrastive translations, differing based on the gender of the translation of *it*: *er*, *sie*, or *es*. The challenge set artificially balances the amount of sentences where *it* is translated to each of these three German pronouns. The appropriate antecedent may be in the main sentence or in a previous sentence. For evaluation, a model needs to produce scores for all three possible translations, which are compared against ContraPro's gold labels.

We create automatic adversarial attacks on ContraPro that modify theoretically inconsequential parts of the context sentence before the occurrence of *it*. Contrary to expectations, we find that accuracy degrades in all adversarial attacks. Results are presented in Figure 1.

5.2 Adversarial Attack Generation

Our three modifications are:

1. **Phrase Addition**: Appending and prepending phrases containing implausible antecedents: The Church is merciful *but that's not the point*. It always welcomes the misguided lamb.

2. **Possessive Extension**: Extending the original antecedent with a possessive noun phrase: I hear ~~her~~ *the doctor's* voice! It resounds to me from heights and chasms a thousand times!

3. **Synonym Replacement**: Replacing the original German antecedent with a synonym of a different gender (note: *der Vorhang* (masc.) and *die Gardine* (fem.) are synonyms meaning *curtain*):
   The curtain rises. It rises. → ~~Der Vorhang~~ *Die Gardine* geht hoch. ~~Er~~ *Sie* geht hoch.

Phrase Addition is applied to all 12,000 ContraPro examples. Depending on suitable conditions, the second and third attack are applied to 3,838 and 1,531 examples, respectively. The Appendix shows results where we vary punctuation and use different added and possessive noun phrases.

5.2.1 Phrase Addition

This attack modifies the previous sentence by appending phrases such as "…*but he wasn't sure*" and also prepending phrases such as "*it is true*:...". A range of other simple phrases can be used, which we leave out for simplicity. In general, all phrases we tried provided lower scores. These attacks introduce a human entity, a pleonastic or an event reference *it* (e.g. *"it is true"*) which are all not plausible antecedents for the anaphoric *it*. We present results for appending "it is true" in Figure 1. Results with using different phrases are presented in the Appendix. In all cases, we prepend or append the same phrase to all ContraPro examples.

5.2.2 Possessive Extension

This attack introduces a new human entity by extending the original antecedent A with a possessive noun phrase e.g., "*the woman's A*". Only two-thirds of the 12,000 ContraPro sentences are linked to an antecedent phrase. Grammar and misannotated antecedents exclude half of the remaining phrases. We put POS-tag constraints on the antecedent phrases before extending them. This reduces our subset to 3,838 modified examples. Our possessive extensions can be humans (*the woman's*), organisations (*the company's*) and names (*Maria's*).

5.2.3 Synonym Replacement

This attack modifies the original German antecedent by replacing it with a German synonym of a different gender. For this we first identify the English antecedent and its most frequent synset in WordNet (Miller, 1995). We obtain a German synonym by mapping this WordNet synsets to GermaNet (Hamp and Feldweg, 1997) synsets. Finally, we modify the correct German pronoun translation to correspond to the gender of the antecedent synonym.

Approximately one quarter of the nouns in our ContraPro examples are found in GermaNet. In 1,531 cases, a synonym of different gender could be identified. Scoring well on the Synonym Replacement attack cannot be done without understanding the pronoun/noun relationship. This attack gets to the core of whether NMT uses CR heuristics instead.

We evaluate a random sample of 100 auto-modified examples as a quality control metric. We note 11 issues with semantically-inappropriate synonyms. Overall, in 14 out of 100 cases, the model switches from correct to incorrect predictions because of synonym-replacement. Only 4 out of these 14 cases come from the questionable synonyms, showing that the drop in ContraPro scores is meaningful.

5.3 Adversarial Attack Results

Our model scores 75.4% on the original ContraPro. This is a very strong result compared to previous work (Müller et al., 2018), largely owing to our model being trained on OpenSubtitles,

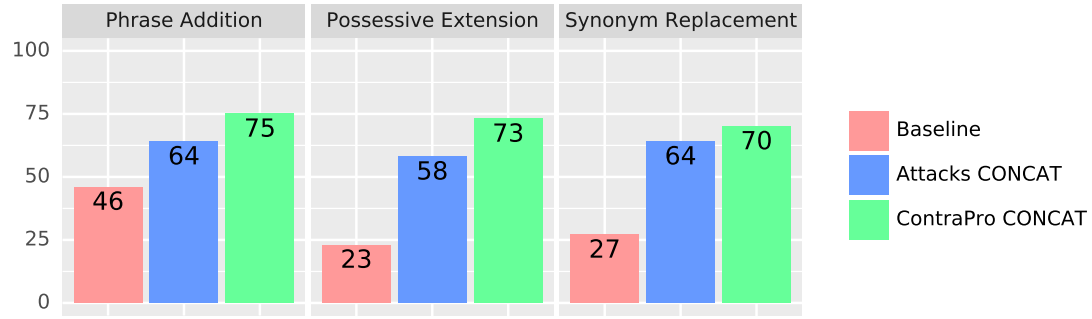


Figure 1: Results with the sentence-level Baseline and CONCAT on ContraPro and three adversarial attacks. The adversarial attacks modify the context, therefore the Baseline model's results on the attacks are unchanged and we omit them. **Phrase**: prepending "it is true: ...". **Possessive**: replacing original antecedent *A* with "Maria's *A*". **Synonym**: replacing the original antecedent with different-gender synonyms. Results for Phrase Addition are computed based on all 12,000 ContraPro examples, while for Possessive Extension and Synonym Replacement we only use the suitable subsets of 3,838 and 1,531 ContraPro examples, respectively.

the same domain as the ContraPro examples. The model scores 72.9% and 69.8% on the ContraPro subsets for the Possessive Extension and Synonym Replacement attacks, respectively. The straightforward adversarial modifications we make drop the ContraPro scores by over 10%, as shown by Figure 1. We analyze examples that are scored incorrectly. Some of the attacks introduce an entity that can in principle be referenced by *it*, like extending the antecedent with "*the company's*". In these cases, the new entity's influence on the model is expected, although ideally, the prediction should not change. More surprisingly, attacks that introduce a human entity drop the scores as well. The two largest examples are appending "*...but he wasn't sure*" and extending the original antecedent with *Maria's*. Our synonym replacement leads to a 6% drop in scores.

Intuitively, the adversarial attacks should not contribute to large drops in scores which is contrary to the empirical evidence. Nevertheless, no attack reduces the model's scores close to the original sentence-level baseline. Thus, we conclude that the concatenation model handles CR, but likely with brittle heuristics. Although the results expose potential issues with the model, it is still difficult to pinpoint the specific problems. This reveals a larger issue with pronoun translation evaluation that cannot be addressed with simple adversarial attacks on existing general-purpose challenge sets. We propose 😸, a more systematic approach that targets each of the previously outlined CR pipeline steps with data synthetically generated from corresponding templates.

6 Templates

Automatic adversarial attacks offer less freedom than templates as many systematic modifications cannot be applied to the average sentence. Thus, our 😸 templates are based on the hypothetical coreference pipeline in Section 3 that target each of the three steps: i) Markable Detection, ii) Coreference Resolution and iii) Language Translation. Our minimalistic templates draw entities from sets of 25 animals, 20 human professions (McCoy et al., 2019), 15 foods, and 5 drinks, along with associated verbs and attributes. We use these sets to fill slots in our templates. Animals and foods are natural choices for subject and object slots referenced by *it*. Restricting our sets to interrelated concepts with generically applicable verbs—all animals eat and drink—ensures semantic plausibility. Other object sets, such as buildings, had more semantic implausibility issues and were not included in the final corpus.

Template Target	Example
Priors	
Grammatical Role	The ***cat*** ate the ***egg***. It (🐱/🥚) was big.
Order	I stood in front of the ***cat*** and the ***dog***. It (🐱/🐶) was big.
Verb	Wow! She unlocked it.
Markable Detection	
Filter Humans	The ***cat*** and the *actress* were happy. However it (🐱) was happier.
Coreference Resolution	
Lexical Overlap	The ***cat*** ate the apple and the *owl* drank the water. It (🐱) ate the apple quickly.
World Knowledge	The ***cat*** ate the *cookie*. It (🐱) was hungry.
Pleonastic it	The *cat* ate the *sausage*. It was raining.
Event Reference	The *cat* ate the *carrot*. It came as a surprise.
Language Translation	
Antecedent Gender	I saw a ***cat***. It(🐱) was big. → Ich habe eine Katze gesehen. Sie (🐱) war groß.

Table 2: Template examples targeting different CR steps and substeps. For German, we create three versions with *er*, *sie*, or *es* as different translations of *it*.

6.1 Template Generation

Our templates consist of a *previous sentence* that introduces at least one entity and a *main sentence* containing the pronoun *it*. We use contrastive evaluation to judge anaphoric pronoun translation accuracy for each template; we create three translated versions for each German gender corresponding to an English sentence, e.g. *"The cat ate the egg. It rained."* and the corresponding *"Die Katze hat das Ei gegessen.* Er/Sie/Es *regnete"*. To fill a template, we only draw pairs of entities with two different genders, i.e. for animal a and food f: $\text{gender}(a) \neq \text{gender}(f)$. This way we can determine whether the model has picked the right antecedent. We refer to "the model picking an antecedent" as the model scoring the target sentence containing the German third person pronoun with the antecedent's gender higher than the provided alternatives.

First, we create templates that analyze priors of the model for choosing a pronoun when no correct translation is obvious. Then, we create templates with correct translations, guided by the three broad coreference steps. Table 2 provides examples for our templates and the results are shown in Figure 2. Template details—entity sets, statistics, etc.—are provided in the Appendix.

6.1.1 Priors

Prior templates do not have a correct answer, but help to understand the model's biases. We expose three priors with our templates: i) grammatical roles prior (e.g. subject) ii) position prior (e.g. first antecedent) and iii) a general prior if no antecedent and only a verb is present.

For i), we create a Grammatical Role template where both subject and object are valid antecedents. We find that in 72.3% of the template instances, the model chooses the object as the antecedent.

For ii), we create a Position template where two objects are enumerated (see Table 2). We create an additional example where the entities order is reversed and test if there are priors for specific nouns or alternatively positions in the sentence.

The model shows a strong prior for neuter by predicting *es* in most cases, even if the two entities are masculine and feminine.

For iii), we create a Verb template, expecting that certain transitive verbs trigger certain object gender choices. We use 100 frequent transitive verbs and create sentences such as the example in Table 2. As expected, *it* is translated to the neuter *es* most of the time, with notable exceptions where the verb is strongly associated with a single noun, e.g. *"Sie hat sie entriegelt"* is scored higher for *"She unlocked it"*. We presume that the reason for this is that *to unlock a door* is very common and door (*Tür*) is feminine in German.

6.1.2 Markable Detection with a Humanness Filter

Before doing the actual CR, the model needs to identify all possible entities that *it* can refer to. We construct a template that contains a human and animal which are in principle plausible antecedents, if not for the condition that *it* does not refer to people. For instance, the model should always choose *cat* in *"The* actress *and the* cat *were hungry. However* it *was hungrier."*. We find that the model instead falls back to translating *it* to the neuter *es* in all cases.

6.1.3 Coreference Resolution

Having determined all possible antecedents, the model has to choose the correct one, relying on semantics, syntax, and discourse. The pronoun *it* can in principle be used as an *anaphoric* (referring to entities), *event reference* or *pleonastic* pronoun (Loáiciga et al., 2017). For the anaphoric *it*, we identify two major ways of identifying the antecedent: lexical overlap and world knowledge. Our templates for these categories are meant to be simple and solvable.

Overlap: Broadly speaking the subject, verb, or object can overlap from the previous sentence to the main sentence, as well as combinations of them. This gives us five templates: i) subject-overlap ii) verb-overlap iii) object-overlap iv) subject-verb-overlap and v) object-verb-overlap. We always use the same template for the context sentence. e.g. *"The* ***cat*** *ate the apple and the* ***owl*** *drank the water."*. For the object-verb-overlap we would then create the main sentence *"It ate the apple quickly."* and expect the model to choose *cat* as antecedent. To keep our overlap templates order-agnostic, we vary the order in the previous sentence by also creating *"The* ***owl*** *drank the water and the* ***cat*** *ate the apple."* However our results in 6.2 show that the model's predictions are almost completely random and are influenced by position priors, e.g., the first mentioned subject, or a prior for the neuter *es* when it needs to decide between the two subjects.

World Knowledge: CR has been traditionally seen as challenging as it requires world knowledge. Our templates test simple forms of world knowledge by using attributes that either apply to animal or food entities, such as *cooked* for food or *hungry* for animals. We then evaluate whether the model chooses e.g. *cat* in *"The* ***cat*** *ate the cookie. It was hungry."* As discussed later, the model occasionally predicts answers that require world knowledge, but most predictions are guided by a prior for choosing the neuter *es* or a prior for the subject.

Pleonastic and Event Templates: For the other two ways of using *it*, event reference and pleonastic-it, we again create a default previous sentence (*"The* ***cat*** *ate the* apple.*")*. For the main sentence, we used four typical pleonastic and event reference phrases such as *"It is a shame"* and *"It came as a surprise"*. We expect the model to correctly choose the neuter *es* as a translation every time and the strong prior for the neuter gender causes the model to do so nearly perfectly.

6.1.4 Translation to German

After CR, the decoder has to translate from English to German. In our contrastive scoring approach the translation of the English antecedent to German is already given. However the decoder is still required to know the gender of the German noun to select between *er, sie* or, *es*. We test this with a list of concrete nouns selected from Brysbaert et al. (2014), which we filter for nouns that occur more than 30 times in the training data. We are left with 2051 nouns which are plugged into the *"I saw a N. It was {big, small}."* template.

6.2 Results

We find that the model performs poorly when actual CR is required. It frequently falls back to choosing the neuter *es* or preferring a position (e.g. first of two entities) for determining the gender. For *Markable Detection* the model always predicts the neuter *es* regardless of the actual genders of the entities.

In the Overlap template, we find that the model fails to recognize the overlap and instead, has a general preference for one of the two clauses. For instance in the case of verb-overlap, the model had a solid accuracy of 64.1% if the verb overlapped from the first clause (*"The cat ate*

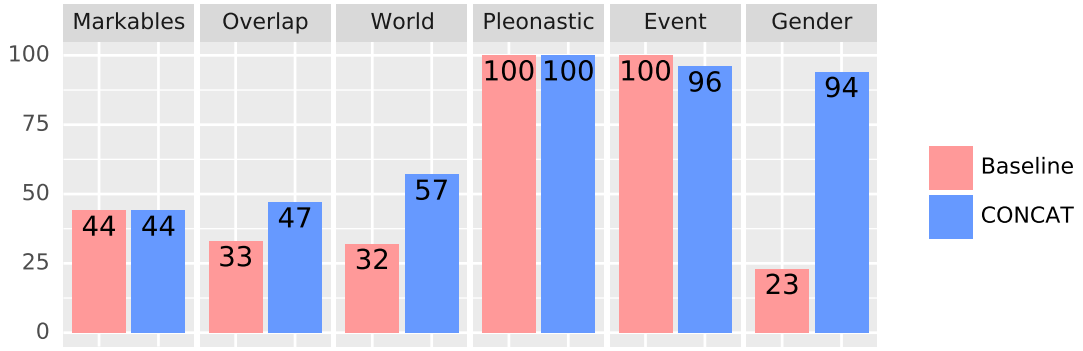


Figure 2: Results comparing the sentence-level baseline to CONCAT on ContraCAT. Pronoun translation pertaining to World Knowledge and language-specific Gender Knowledge benefits the most from additional context.

and the dog drank. It ate a lot.") but a weak accuracy of 39.0% when the verb overlapped from the second clause (*"The cat ate and the dog drank. It drank a lot."*.) The overall accuracy for the overlap templates is 47.2%, with little variation across the types of overlap. Adding more overlap, e.g., by overlapping both the verb and object (*"It ate the apple happily"*), yields no improvement. Overall, the model pays very little attention to overlaps when resolving pronouns.

We also see weak performance for world knowledge. An accuracy of 55.7% is slightly above the heuristic of randomly choosing an entity (= 50.0%). With a strong bias for the neuter *es*, the model has a high accuracy of 96.2% for event reference and pleonastic templates, where *es* is always the correct answer. Based on the strong performance on the Gender template in 6.1.4, we conclude the model consistently memorized the gender of concrete nouns. Hence, CR mistakes stem from Step 1 or Step 2, suggesting that the model failed to learn proper CR.

7 Augmentation

We present an approach for augmenting the training data. While challenging for NLP, we focus on a narrow problem which lends itself to easier data manipulation. Our previous analyses show that our model is capable of modeling the gender of nouns. However, they also show a strong prior to translate *it* to *es* and very little CR capability. Our goal with the augmentation is to break off the strong prior and test if this can give rise to better CR in the model.

We attempt to do this by augmenting our training data and call it Antecedent-free augmentation (AFA). We identify candidates for augmentation as sentences where a coreferential *it* refers to an antecedent not present in the current or previous sentence (e.g., *I told you before. <SEP> It is red. → Ich habe dir schonmal gesagt. <SEP> Es ist rot.*). We create augmentations by adding two new training examples where the gender of the German translation of "it" is modified (e.g., the two new targets are "*Ich habe dir schonmal gesagt. <SEP> Er ist rot.*" and "*Ich habe dir schonmal gesagt. <SEP> Sie ist rot.*"). The source side remains the same. An additional example is shown in Table 3. Antecedents and coreferential pronouns are identified using a CR tool (Clark and Manning, 2016a; Clark and Manning, 2016b). We fine-tune our already trained concatenation model on a dataset consisting of the candidates and the augmented samples. As a baseline, we fine-tune on the candidates only so as to confidently say that any potential improvements come from the augmentations.

7.1 Results

7.1.1 Adversarial Attacks

AFA provides large improvements, scoring 85.3% on ContraPro. Results are shown in Figure 3. The AFA baseline (fine-tuning on the augmentation candidates only) improves by 1.94%,

Antecedent-free augmentation	
Source	You let me worry about that. <SEP> How much you take for it?
Reference	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet er?
Augmentation 1	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet sie?
Augmentation 2	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet es?

Table 3: Examples of training data augmentations. The source side of the augmented examples remains the same.

presumably because many candidates consist of coreference chains of "it" and the model learns they are important for coreferential pronouns. However, the improvement is small compared to AFA.

Results on ContraPro for each gender (see Appendix) show that performance on *er* and *sie* is substantially increased, suggesting that the augmentation successfully removes the strong bias towards *es*. Templates provide further evidence about this. Although, the adversarial attacks lower AFA scores, in contrast to CONCAT, the model is more robust and the performance degradation is substantially lower (except on the synonym attack). We experimented with different learning rates during fine-tuning and present results with the LR that obtained the best baseline ContraPro score. Detailed scores in the Appendix show how LR can balance the scores across the three different genders. Furthermore, CONCAT and AFA obtain 31.5 and 32.2 BLEU on ContraPro, respectively, showing that this fine-tuning procedure, which is tailored to pronoun translation, does not lead to any degradation in general translation quality.

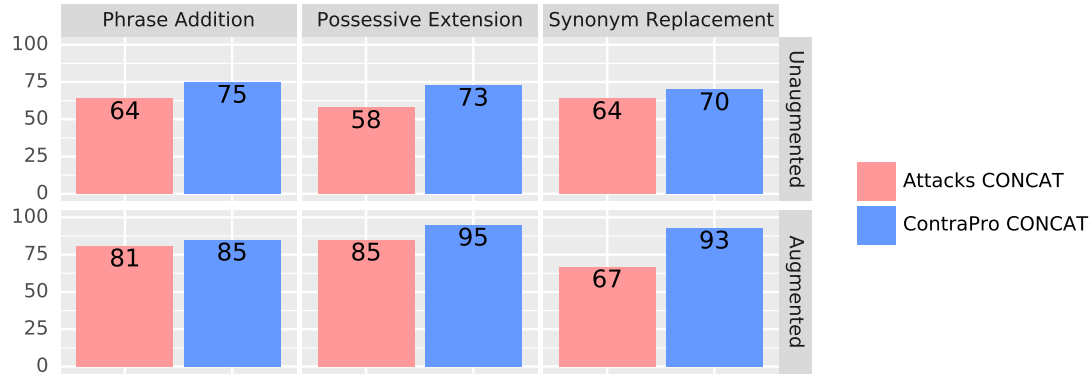


Figure 3: Results comparing unaugmented and augmented CONCAT on ContraPro and same 3 attacks as in Figure 1. Results with non-augmented CONCAT are the same as Figure 1.

7.1.2 Templates

From the prior templates, we observe that the prior over gender pronouns is more evenly spread and not concentrated on *es*. This also provides for a more even distribution on the Position and Role Prior template. The results on the prior templates are presented in the Appendix. The augmented model is also substantially better on markable detection, improving by 27.6%. Results for templates are presented in Figure 4.

No improvements are observed on the World Knowledge template. Pleonastic cases are still reasonably handled, although not perfectly as with CONCAT. The Event template identifies a systematic issue with our augmentation. We presume this is as a result of the CR tool marking cases where *it* refers to events. We do not apply any filtering and augment these cases as well, thus create wrong examples (an event reference *it* cannot be translated to *er* or *sie*). As a result, the scores are significantly lower compared to CONCAT. We note that this issue with our model is not visible on ContraPro and the adversarial attacks results. In contrast, the Event template

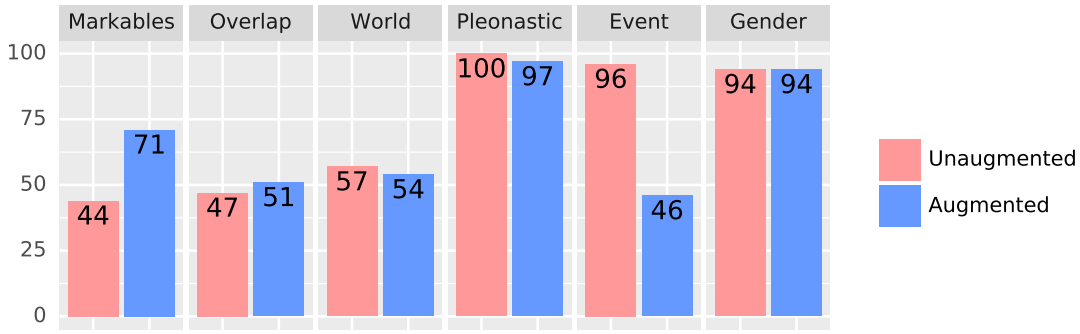


Figure 4: ContraCAT results with unaugmented and augmented CONCAT. We speculate that readjusting the prior over genders in augmented CONCAT explains the improvements on Markable and Overlap.

easily identifies this problem.

AFA performs on par with the unaugmented baseline on the Gender template. However, despite increasing by 3.8%, results on Overlap are still underwhelming. Our analysis shows that augmentation helps in changing the prior. We believe this provides for improved CR heuristics which in turn provide for an improvement in coreferential pronoun translation. Nevertheless, the Overlap template shows that augmented models still do not solve CR in a fundamental way.

8 Conclusion

In this work, we study how and to what extent CR is handled in context-aware NMT. We show that standard challenge sets can easily be manipulated with adversarial attacks that cause dramatic drops in performance, suggesting that NMT uses a set of heuristics to solve the complex task of CR. Attempting to diagnose the underlying reasons for these results, we propose targeted templates which systematically test the different aspects necessary for CR. This analysis shows that while some type of CR such as pleonastic and event CR are handled well, NMT does not solve the task in an abstract sense. We also propose a data augmentation approach which substantially improves performance on some metrics, but it does not change the general conclusions we infer from the templates. Future work should be evaluated on our adversarial attacks and ContraCAT, which we publicly release, to realistically estimate the ability of NMT to robustly do CR.

Acknowledgments

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement № 640550). This work was also supported by DFG (grant FR 2829/4-1). We thank Alexandra Chronopoulou for the valuable comments and helpful feedback.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas. *Behavior Research Methods*, 46:904–911.

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, Austin, Texas, November. Association for Computational Linguistics.

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, Berlin, Germany, August. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2018. Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium, October-November. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

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. 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).

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*, December.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China, November. Association for Computational Linguistics.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2006. Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding. In *Final Report of the 2006 JHU Summer Workshop*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the 15th conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? Disambiguating the Different Readings of the Pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark, September. Association for Computational Linguistics.

António Lopes, M Amin Farajian, Rachel Bawden, Michael Zhang, and André Martins. 2020. Document-level Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September. Association for Computational Linguistics.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *WMT 2018*, Brussels, Belgium. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Dario Stojanovski 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: Research Papers*, pages 49–60, Brussels, Belgium, October. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019. Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland, August. European Association for Machine Translation.

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.

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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

Xiaosen Wang, Hao Jin, and Kun He. 2019. Natural Language Adversarial Attacks and Defenses in Word Level. *arXiv preprint arXiv:1909.06723*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A Preprocessing Details and Model Hyper-Parameters

We use OpenSubtitles2018[2] (Lison and Tiedemann, 2016) as training data. We tokenize the dataset using the Moses scripts[3] (Koehn et al., 2006). We BPE-split the data by jointly computing them on English and German using 32K merge operations. We remove all samples where the main sentence exceeds 100 tokens on the BPE-level or the concatenated sample contains more than 200 tokens. ContraPro is built using OpenSubtitles and contains samples from it. As a result, we remove the entire documents from which the ContraPro samples originate from in order to remove any exact duplicates from ContraPro or similar contexts which may lead to unfair advantages for the models. This still leaves some exact duplicates between our training data and ContraPro which we also remove. The model is finally trained on ≈16.7M samples.

We train the transformer models with a batch size of 4096. We use an initial learning rate of 10^{-4} and we lower it by a factor of 0.7 if there are no improvements on the validation perplexity for 8 checkpoints. We save a checkpoint every 4000 updates.

The transformer models we use are a 6 layer encoder/decoder with 8 attention heads. The model size is 512 and the size of the feed-forward layers is 2048. We tie the source, target and output embeddings. We use label smoothing with 0.1 and dropout in the transformer of 0.1. Models are trained on 2 GTX 1080 GPUs with 8GB RAM. The final model is an average of the 8 best checkpoints based on validation perplexity. The models we train are implemented in Sockeye (Hieber et al., 2017).

B Complete List of Automatic Attacks on ContraPro

B.1 Adding Phrases

As a reference to the scores shown in Table 4, our model has a score of 0.754 on unmodified ContraPro. *C* is the original context sentence from ContraPro.

B.2 Possessive Extension

These were applied to 3,838 ContraPro examples. As a reference to scores shown in Table 5, our model has a score of 72.9% on the unmodified subset of ContraPro. *A* refers to the original antecedent noun phrase.

[2] `http://opus.nlpl.eu/OpenSubtitles-v2018.php`
[3] `https://github.com/moses-smt/mosesdecoder`

Modification	ContraPro Score
he/she said: "*C*"	66.5 / 66.3
it is true:"*C* / it is true: *C*	55.2 / 63.5
C and it is true. / *C*. it is true. / *C* and that is true.	65.1 / 57.9 / 70.4
C but he wasn't sure. / *C*. but he wasn't sure.	69.0 / 65.8
C but that's not the point. / *C*. but that's not the point.	70.7 / 67.5
C but there is a catch. /*C*. but there is a catch.	67.4 / 64.6
C but why. / *C*. but why.	74.0 / 71.5

Table 4: Scores for each *Adding-Phrase*-modification. Slightly altered modifications are indicated with "/"

Modification	ContraPro Score
... the woman's *A* ...	63.4
... the man's *A* ...	66.5
... my mother's *A* ...	70.9
... my father's *A* ...	72.2
... the dog's *A* ...	66.8
... the cat's *A* ...	67.4
... the doctor's (vom Arzt/von der Ärztin) *A* ...	66.7 / 66.4
... *A* of my best friend's mother ...	60.0
... the government's *A* ...	68.5
... the company's *A* ...	63.0
... Maria's *A* ...	58.3
... Lisa's *A* ...	60.3
... Bolsena's *A* ...	60.3
... Peter's *A* ...	59.0
... Robert's *A* ...	60.5
... David's *A* ...	60.6

Table 5: Scores for each *Possessive-Extension*-modification. For German we append the possessive noun phrase with "von"(=of).

B.3 Synonym replacement

These were applied to 1,531 ContraPro examples. Our model has a score of 69.8% on the unmodified subset of ContraPro. When replacing with different-gender synonyms we drop to a score of 64.1%.

C Template Generation

C.1 Vocabulary

Our templates draw from the sets of entities shown in Table 6 and Table 7. The translations in German are shown in brackets. We note that all entities appear in the training dataset we use to train our models. The least frequent entity ("kangaroo") appears 134 times.

We also use four event- and pleonastic-it phrases which are used as the main sentence in the templates and referred to later.

Event: It came as a surprise (Es kam überraschend), It actually happened (Es ist tatsächlich passiert), It resulted in chaos (Es führte zu Chaos), It was a funny situation (Es war eine lustige Situation)

Pleonastic: It was raining (Es regnete), It is a shame (Es ist eine Schande), It seemed this was unnecessary (Es schien, dass dies unnötig war), It is hard to believe this is true (Es ist schwer zu glauben , dass das wahr ist)

ANIMALS		PROFESSIONS
dog (Hund)	giraffe (Giraffe)	professor (Professor(in))
wolf (Wolf)	mouse (Maus)	student (Student(in))
bear (Bär)	duck (Ente)	judge (Richter(in))
tiger (Tiger)	turtle (Schildkröte)	secretary (Sekretär(in))
lion (Löwe)	owl (Eule)	doctor (Arzt/Ärztin)
rabbit (Hase)	dove (Taube)	lawyer (Anwalt/Anwältin)
monkey (Affe)	goat (Ziege)	scientist (Wissenschaftler(in))
eagle (Adler)	sheep (Schaf)	manager (Manager(in))
frog (Frosch)	squirrel (Eichhörnchen)	artist (Künstler(in))
cat (Katze)	horse (Pferd)	actor (Schauspieler)
cow (Kuh)	pig (Schwein)	actress (Schauspielerin)
zebra (Zebra)	kangaroo (Känguru)	
deer (Reh)		

Table 6: Vocabulary of entities used in templates.

FOOD		DRINKS	
cookie (Keks)	cake (Kuchen)	tea (Tee)	juice (Saft)
carrot (Karotte)	hot dog (Hotdog)	milk (Milch)	lemonade (Limonade)
cheese (Käse)	apple (Apfel)	water (Wasser)	
nut (Nuss)	fruit (Frucht)		
sausage (Wurst)	pizza (Pizza)		
bread (Brot)	egg (Ei)		
meat (Fleisch)	ice cream (Eis)		
steak (Steak)			

Table 7: Vocabulary of entites used in templates.

C.2 Template Statistics

For each template, we report the number of lines it contains in Table 8.

Template	Number of lines
Grammatical Role Prior	1000
Position Prior	828
Verb Prior	600
Markable Detection (animacy)	2560
Verb Overlap	2240
Object Overlap	5376
Subject Overlap	4992
Object-Verb Overlap	5376
Subject-Verb Overlap	4992
World Knowledge	2500
Event	1500
Pleonastic	1500
Gender	4102

Table 8: Number of test sentences for each template.

C.3 Template Definitions

The template definitions are shown in Table 9. We refer to animals with A, professions as P, food as F, drinks as D. When creating a concrete animal, food or drink X_i, we use the definite

article "the" ("der/die/das" in German). On the German side, we underline the options that we give the model for the three German genders.

Template	English definition	German definition
Grammatical Role Prior	A ate F. It was {big, small, large, tiny}.	A hat F gegessen. Er/Sie/Es war {groß, klein, riesig, winzig}.
Position Prior	I stood in front of A_i and A_j. It was {big, small, large, tiny}.	Ich stand vor A_i und A_j. Er/Sie/Es war {groß, klein, riesig, winzig}.
Verb Prior	Wow! I/You/He/She/We/They $V_{past+transitive}$ it.	Wow! Ich/Du/Er/Sie/Wir/Sie haben er/sie/es $V_{past+transitive}$.
Markable Detection (filter humans)*	A and P were {hungry, tired, happy, nice}. However it was {hungrier, more tired, happier, nicer}.	A und P waren {hungrig, müde, glücklich, nett}. Aber er/sie/es war {hungriger, müder, glücklicher, netter}.
Verb Overlap*	A_i {ate, drank} and A_j {ate, drank}. It {ate, drank} {a lot, quickly, slowly happily}.	A_i hat {gegessen, getrunken} und A_j hat {gegessen, getrunken}. Er/Sie/Es hat {viel, schnell, langsam, fröhlich} {gegessen, getrunken}.
Object Overlap*	A_i ate F and A_j drank D. It liked {F, D}.	A_i hat F gegessen und A_j hat D getrunken. Er/Sie/Es mochte {F, D}.
Subject Overlap*	A_i ate F and A_j drank D. {A_i, A_j} liked it.	A_i hat F gegessen und A_j hat D getrunken. {A_i, A_j} mochte ihn/sie/es.
Object-Verb Overlap*	A_i ate F and A_j drank D. It {ate F, drank D} quickly.	A_i hat F gegessen und A_j hat D getrunken. Er/Sie/Es hat {F schnell gegessen, D schnell getrunken}.
Subject-Verb Overlap*	A_i ate F and A_j drank D. {A_i ate, A_j drank} it quickly.	A_i hat F gegessen und A_j hat D getrunken. {A_i, A_j} hat ihn/sie/es schnell {gegessen, getrunken}.
World Knowledge	A ate F. It {was hungry, was looking around, was running around, was tired, was happy} / {had a sweet/bitter/sour taste, was cooked, had gone bad}.	A hat F gegessen. Er/Sie/Es {war hungrig, schaute sich um, rannte herum, war müde, war glücklich} / {hatte einen süßen/bitteren/sauren Geschmack, war gekocht, war schlecht geworden}.
Event	A ate F. EVENT-PHRASE	A hat F gegessen. EVENT-PHRASE.
Pleonastic	A ate F. PLEONASTIC-PHRASE	A hat F gegessen. PLEONASTIC-PHRASE.
Gender	I saw a $N_{concrete}$. It was {big, small}.	Ich sah ein/eine/einen $N_{concrete}$. Er/Sie/Es war {groß, klein}.

Table 9: Template definitions. * We switch the position (first or second) of the two involved entities E_i and E_j.

C.4 Prior Results

For the templates that do have a correct answer, we show results in the main paper. In Table 10, Table 11 and Table 12 we show the results on the grammatical, position and verb prior templates.

Model	subject	object
CONCAT	20.7%	72.3%
AFA	52.2%	47.8%

Table 10: Grammatical Role template for testing prior of choosing subject or object as antecedent to translate *it*. If numbers do not add up to 100%, it is because the model chose neither the subject nor object. This is usually the neuter *es*.

Model	first	second	same antecedent
CONCAT	0.0%	3.1%	60.8%
AFA	0.2%	13.0%	74.9%

Table 11: Position template for testing prior for first or second enumerated object as antecedent to translate *it*. If numbers do not add up to 100%, it is because the model chose neither the first nor second object. This is usually the neuter *es*.

Model	masculine	feminine	neuter
CONCAT	5.7%	2.8%	91.5%
AFA	43.7%	27.5%	28.8%

Table 12: Verb template for testing prior for the three genders, only conditioned on a transitive verb.

D Augmentation

D.1 Details

For all augmentations we use Spacy's dependency parser[4] in order to determine the case of the pronoun. This is necessary because the feminine ("sie") and neuter ("es") pronoun are the same in nominative and accusative, but the masculine is not ("er" and "ihn"). We fine-tuned on 207K for the antecedent-free augmentations.

D.2 Fine-Tuning Learning Rate Analysis

We conducted 3 different fine-tuning experiments where we varied the learning rate. We used a learning rate of $2 * 10^{-6}$, $2 * 10^{-7}$ and $2 * 10^{-8}$. The initial concatenation model was trained with an initial LR of $2 * 10^{-4}$ and when it converged, the learning rate was $7.82 * 10^{-8}$. Results are presented in Table 13. As before, we average 8 checkpoints before evaluating our models.

	total	er	sie	es
CONCAT	75.4	64.0	66.8	95.3
AFA lr=10^{-6}	78.4	81.0	81.9	72.5
AFA lr=10^{-7}	85.3	88.2	90.6	77.2
AFA lr=10^{-8}	81.3	73.9	77.3	92.7

Table 13: Challenge set performance for each pronoun.

[4] https://spacy.io/usage/linguistic-features#dependency-parse

As the goal with the augmentations is to remove the strong bias towards neuter, we only evaluate the different LR models on the "er", "sie" and "es" accuracy on ContraPro. Performance on "er" and "sie" improves in all experiments, but it improves by far the most using a LR of $2 * 10^{-7}$. Performance on "es" gets worse as the LR increases. However, very low LR also does not provide for large improvements on "er" and "sie". We show that the LR is an important hyper-parameter in order to balance the performance on all pronouns. Admittedly, one may opt for a lower learning rate because, as the training data shows, "it" tends to be translated to "es", so it is undesirable to significantly drop performance on "es" because in practice these errors will be more visible.

Chapter 7

Conclusion

7.1 Summary

In this thesis we present our work on context-aware neural machine translation. We give a detailed overview of the broad machine learning and natural language processing models and methods our work is based on. Furthermore, we provide a detailed analysis of previous work on context-aware neural machine translation and the relevant discourse-level phenomena in MT.

In Chapter 2, we present a method that can be used to determine the importance of different discourse-level phenomena using automatically-created oracle signals. We compare different RNN- and Transformer-based context-aware models and conclude that both coreference resolution and coherence are important for better translation quality.

In Chapter 3, we present a curriculum learning method for better anaphoric pronoun translation in MT which is based on the work in Chapter 2. We show that training context-aware models in a manner similar to how humans learn can provide for improvements in some limited experimental setups.

In Chapter 4, we present a previously unexplored area and propose to evaluate context-aware NMT models in multi-domain setups and on domains not seen during training. Furthermore, we propose two context-aware NMT models capable of handling large context and show that context is helpful in determining domain in the aforementioned scenarios.

In Chapter 5, we present our work on combining models that encode local and global context in a fine- and coarse-grained way, respectively. We show that the improvements obtained from such models are complimentary and that this is a

promising research idea to further pursue.

Finally, in Chapter 6, we show that adversarially attacking existing challenge sets for anaphoric pronoun translation evaluation can provide for different results and conclusions about the abilities of context-aware models. As a result, we propose a template test set named ContraCAT that evaluates specifics steps of a hypothetical coreference resolution pipeline in MT. Finally, we propose a data augmentation technique to deal with the potential bias in MT models with regards to pronoun translation. Results on existing challenge sets and our adversarial attacks show that pronoun translation is improved by using this technique, but the method does not consistently improve on ContraCAT, particularly in the cases where strong reasoning is required.

7.2 Avenues for Improvement

The work presented in this thesis, proposed new methods for analysis of discourse-level phenomena in MT (Chapter 2), new methods for better coreference resolution in MT (Chapter 3), novel models and evaluation setups (Chapter 4) and conducted a detailed analysis of coreference resolution (Chapter 6). Despite the thorough work and analysis in each chapter, the work contains some shortcomings.

In Chapter 2, we proposed using oracle signals to determine the importance of different discourse-level phenomena in MT. In this work, we only considered coreference resolution and coherence. However, we took a more narrow view of coherence and only considered the aspect of consistency to surface formulations. More precisely, we only looked at repeated words. Furthermore, this work did not consider other discourse-level phenomena. It is important to point out that the oracle signals we used were automatically determined and they can be adapted to other languages and discourse-level phenomena with relative ease.

In Chapter 3, we proposed a curriculum learning method for improving anaphora resolution in MT and we showed that the method obtains promising results. However, in our experimental setup, the curriculum learning method only provided robust improvements in a limited setting where we used a certain learning rate. Nevertheless, our results showed that using this specific learning rate provides a compromise between pronoun and general translation quality.

In Chapter 4, we argued that context-aware MT models should be evaluated in multi-domain setups. Additionally, we provided two novel context-aware models. Although the models we proposed were novel at the time, they may be seen as

relatively simple. However, we use relatively simple models in order to be able to make clear conclusions about whether domain information is being encoded in our models.

In Chapter 5, we laid the groundwork for our hypothesis that local and global context should be modeled in different ways. This distinction has not been made in previous work. However, the combination of local and global context in the work of Chapter 5 was done in a relatively simple way and mainly consisted of combining existing methods. Further work will be necessary to precisely define the distinction between local and global context and new models are likely to be necessary to address this problem in a more principled way.

In Chapter 6, we presented work that showed that existing challenge sets for evaluating anaphoric pronoun translation can be manipulated with adversarial attacks. Furthermore, we presented a novel template test set for coreference resolution evaluation in MT. We carefully designed the adversarial attacks in our experiments and made a strong effort for the modifications we made to be compatible with the original text. However, an argument can be made against this approach as it artificially modifies the test set. However, an NMT model should be able to handle any type of text, especially considering that our modifications were made so as to not affect any aspect related to pronoun translation. This points to the brittleness of current NMT models. Furthermore, both our adversarial attacks and ContraCAT point to that NMT models use heuristics to solve the problem. However, our work does not precisely identify what all of those heuristics are and further work is necessary to fully address this problem.

7.3 Future Work

Recent work in context-aware NMT has made tremendous progress. Nonetheless, there is still large room for improvement across several axes. One way future work can provide for better context-aware NMT models is to improve the way large context is handled. Several works have attempted to use large context, but done so in a way that requires fine-grained access to all tokens. As already argued in this thesis, future work may focus on ways to address the issue of large context by modeling it in a more coarse-grained way that does not require one to deal with backpropagation across enormous distances.

Another potential issue with current context-aware NMT approaches is that they are trained in a straightforward way with the only difference from sentence-level models being that contextual information is used. However, it is reasonable

to assume that useful contextual signals are very rare, compared to the useful information originating from the sentence being translated. Considering how current machine learning models are trained, it may be possible that current context-aware models cannot properly learn how and when to use contextual information. We believe that this can be remedied by using minimum risk training or reinforcement learning with discourse-specific rewards. One can imagine training context-aware NMT models with specific signals reward the models for better anaphoric pronoun translation, better coherence and so on. Similar work has been done by Jauregi Unanue et al. (2020). It may be interesting to explore the utility of meta-learning as a method for better few-shot learning. One can imagine fine-tuning context-aware models with meta-learning on training samples predetermined as having useful contextual information and therefore emphasizing the importance of context.

Finally, as already pointed out in Chapter 6, current discourse-specific evaluation methods are in need of improvement. In this thesis, we showed this for coreference resolution, but similar works will be necessary for other discourse-level phenomena. Furthermore, an open question is to what extent should our methods for evaluation be based on scoring translation pairs. While this is a convenient method for automatic evaluation, the results it provides may not fully correspond to how the model behaves when it is tasked with free translation.

We like to conclude by emphasizing the importance of context-aware machine translation. While several challenges still lie ahead in the field, mostly pertaining to better modeling and evaluation, this problem is of high importance. The MT community strives for human-level translation performance and using contextual information is necessary to achieving this goal.

Bibliography

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Ankur Bapna and Orhan Firat. Non-Parametric Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1921–1931, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1191. URL https://www.aclweb.org/anthology/N19-1191.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://www.aclweb.org/anthology/N18-1118.

V. Becher. Explicitation and implicitation in translation. 2011.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

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

D. Blakemore. *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge Studies in Linguistics. Cambridge University Press, 2002. ISBN 9781139437301. URL https://books.google.co.ls/books?id=9142DhlPifUC.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://www.aclweb.org/anthology/W16-2301.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

Xinyi Cai and Deyi Xiong. A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.iwdp-1.3.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1053. URL https://www.aclweb.org/anthology/D16-1053.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

BIBLIOGRAPHY

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. Towards coherent and cohesive long-form text generation. In *Proceedings of the First Workshop on Narrative Understanding*, pages 1–11, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2401. URL https://www.aclweb.org/anthology/W19-2401.

Kevin Clark and Christopher D. Manning. 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, 2016a. doi: 10.18653/v1/P16-1061. URL http://www.aclweb.org/anthology/P16-1061.

Kevin Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, 2016b. doi: 10.18653/v1/D16-1245. URL http://www.aclweb.org/anthology/D16-1245.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://www.aclweb.org/anthology/P19-1285.

R. De Beaugrande, W.U. Dressler, and Green & Co Longmans. *Introduction to Text Linguistics*. Lll Series. Longman, 1981. ISBN 9780582554863. URL https://books.google.mk/books?id=mvJsAAAAIAAJ.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 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, 2013. URL http://aclweb.org/anthology/N13-1073.

Peter W. Foltz, Walter Kintsch, and Thomas K Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25

(2-3):285–307, 1998. doi: 10.1080/01638539809545029. URL https://doi.org/10.1080/01638539809545029.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/gehring17a.html.

Zhengxian Gong, Min Zhang, and Guodong Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2504. URL https://www.aclweb.org/anthology/W15-2504.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Liane Guillou and Christian Hardmeier. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1100.

Liane Guillou and Christian Hardmeier. Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1513. URL https://www.aclweb.org/anthology/D18-1513.

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

Jan Hajic. Ruslan-an mt system between closely belated languages. In *Third Conference of the European Chapter of the Association for Computational Linguistics*, 1987.

M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

Christian Hardmeier. 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), 2012.

Christian Hardmeier. *Discourse in statistical machine translation*. PhD thesis, Acta Universitatis Upsaliensis, 2014.

Christian Hardmeier and Marcello Federico. 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, 2010.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 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, 2013.

Hany Hassan, Anthony Aue, C. Chen, Vishal Chowdhary, J. Clark, C. Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, W. Lewis, M. Li, Shujie Liu, T. Liu, Renqian Luo, Arul Menezes, Tao Qin, F. Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and M. Zhou. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Donald Olding Hebb. *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall, 1949.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 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 2018. Association for Machine Translation in the Americas. URL `https://www.aclweb.org/anthology/W18-1820`.

Geoffrey E Hinton. Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10):428–434, 2007.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Inigo Jauregi Unanue, Nazanin Esmaili, Gholamreza Haffari, and Massimo Piccardi. Leveraging discourse rewards for document-level neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4467–4482, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.395.

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

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

Rod Johnson, Maghi King, and Louis des Tombe. Eurotra: A multilingual system under development. *Computational Linguistics*, 11(2-3):155–169, 1985.

Marcin Junczys-Dowmunt. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5321. URL https://www.aclweb.org/anthology/W19-5321.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1294. URL https://www.aclweb.org/anthology/D19-1294.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online,

November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.175. URL https://www.aclweb.org/anthology/2020.emnlp-main.175.

Yunsu Kim, Thanh Tran, and Hermann Ney. When and why is document-level context useful in neural machine translation? In *DiscoMT@EMNLP*, 2019.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Catherine Kobus, Josep Crego, and Jean Senellart. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378. INCOMA Ltd., 2017. doi: 10.26615/978-954-452-049-6_049. URL https://doi.org/10.26615/978-954-452-049-6_049.

Tom Kocmi and Ondřej Bojar. 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., 2017. doi: 10.26615/978-954-452-049-6_050. URL https://doi.org/10.26615/978-954-452-049-6_050.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Philipp Koehn and Christof Monz, editors. *Proceedings on the Workshop on Statistical Machine Translation*, New York City, June 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-3100.

Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133, 2003. URL https://www.aclweb.org/anthology/N03-1017.

Sachith Sri Ram Kothur, Rebecca Knowles, and Philipp Koehn. Document-Level Adaptation for Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-2708.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 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 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1050.

Mirella Lapata and Regina Barzilay. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 1085–1090, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4810. URL https://www.aclweb.org/anthology/W17-4810.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://www.aclweb.org/anthology/D18-1512.

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

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.322. URL https://www.aclweb.org/anthology/2020.acl-main.322.

Grace W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14:29, 2020. ISSN 1662-5188. doi: 10.3389/fncom.2020.00029. URL https://www.frontiersin.org/article/10.3389/fncom.2020.00029.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1147.

Frederick Liu, Han Lu, and Graham Neubig. 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, 2018. URL http://aclweb.org/anthology/N18-1121.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. Findings of the 2017 discomt shared task on cross-lingual pronoun prediction. In *The Third Workshop on Discourse in Machine Translation*, 2017.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Lin-

guistics. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166.

Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.321. URL https://www.aclweb.org/anthology/2020.acl-main.321.

Valentin Macé and Christophe Servan. Using whole document context in neural machine translation. *arXiv preprint arXiv:1910.07481*, 2019.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/martins16.html.

Sameen Maruf and Gholamreza Haffari. 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, 2018. URL http://aclweb.org/anthology/P18-1118.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313. URL https://www.aclweb.org/anthology/N19-1313.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level machine translation: Methods and evaluation. *ArXiv*, abs/1912.08494, 2019b.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://www.aclweb.org/anthology/2020.acl-main.448.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, page 10, October 2012.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 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, 2018. URL http://aclweb.org/anthology/D18-1325.

Lesly Miculicich Werlen and Andrei Popescu-Belis. Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4802. URL https://www.aclweb.org/anthology/W17-4802.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

George A Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. Effective use of target-side context for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4483–4494, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.396.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 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 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-6307.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

A. Popescu-Belis. Context in neural machine translation: A review of models and evaluations. *ArXiv*, abs/1901.09115, 2019.

Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong, editors. *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-6500.

Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-6319.

Ofir Press and Lior Wolf. 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, 2017. URL http://aclweb.org/anthology/E17-2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September 2018. doi: 10.1162/coli_a_00322. URL https://www.aclweb.org/anthology/J18-3002.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.693. URL https://www.aclweb.org/anthology/2020.acl-main.693.

Rico Sennrich. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2060.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL https://www.aclweb.org/anthology/N16-1005.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, 2016b. doi: 10.18653/v1/P16-1162. URL http://aclweb.org/anthology/P16-1162.

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. 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, 2017. URL http://aclweb.org/anthology/E17-3017.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784, 2016.

Karin M Sim Smith. *Coherence in Machine Translation*. PhD thesis, University of Sheffield, 2018.

Karin Sim Smith and Lucia Specia. Assessing crosslingual discourse relations in machine translation. *ArXiv*, abs/1810.03148, 2018.

Margita Šoštarić, Christian Hardmeier, and Sara Stymne. Discourse-related language contrasts in English-Croatian human and machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 36–48, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6305. URL https://www.aclweb.org/anthology/W18-6305.

Dario Stojanovski and Alexander Fraser. 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 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-6306.

Dario Stojanovski and Alexander Fraser. Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland, 19–23 August 2019a. European Association for Machine Translation. URL https://www.aclweb.org/anthology/W19-6614.

Dario Stojanovski and Alexander Fraser. Combining local and document-level context: The lmu munich neural machine translation system at wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared*

Task Papers, Day 1), pages 400–406, Florence, Italy, August 2019b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W19-5345.

Dario Stojanovski and Alexander Fraser. Addressing zero-resource domains using document-level context in neural machine translation. *arXiv preprint arXiv:2004.14927*, 2020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1168. URL https://www.aclweb.org/anthology/D19-1168.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, 2017. URL http://aclweb.org/anthology/W17-4811.

Yiqi Tong, Jiangbin Zheng, Hongkang Zhu, Yidong Chen, and Xiaodong Shi. A document-level neural machine translation model with dynamic caching guided by theme-rheme information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4385–4395, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.388.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6312. URL https://www.aclweb.org/anthology/W18-6312.

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

Ferhan Ture, Douglas W Oard, and Philip Resnik. 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, 2012.

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

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 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, 2018. URL http://aclweb.org/anthology/P18-1117.

Elena Voita, Rico Sennrich, and Ivan Titov. Context-Aware Monolingual Repair for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 876–885, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1081. URL https://www.aclweb.org/anthology/D19-1081.

Elena Voita, Rico Sennrich, and Ivan Titov. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy,

July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://www.aclweb.org/anthology/P19-1116.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, 2017. URL http://aclweb.org/anthology/D17-1301.

Rui Wang, Masao Utiyama, and Eiichiro Sumita. 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, 2018. URL http://aclweb.org/anthology/P18-2048.

Tian Wang and Kyunghyun Cho. Larger-context language modelling. *arXiv preprint arXiv:1511.03729*, 2015.

Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. *Proceedings of the Workshop on Discourse in Machine Translation*, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3300.

Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier, editors. *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-25. URL https://www.aclweb.org/anthology/W15-2500.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors. *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-48. URL https://www.aclweb.org/anthology/W17-4800.

Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1097.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.530. URL https://www.aclweb.org/anthology/2020.acl-main.530.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7338–7345, Jul. 2019. doi: 10.1609/aaai.v33i01.33017338. URL https://ojs.aaai.org/index.php/AAAI/article/view/4721.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. Enhancing Context Modeling with a Query-Guided Capsule Network for Document-level Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1164. URL https://www.aclweb.org/anthology/D19-1164.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc., 2019b.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 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, 2018a. URL http://aclweb.org/anthology/D18-1049.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online,

November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.81. URL https://www.aclweb.org/anthology/2020.emnlp-main.81.

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