Karlsruhe Institute of Technology

Document-Level Pretraining for Neural Chat Translation

Bachelor's Thesis of

Tobias Kaiser

at the Department of Informatics
Institute for Anthropomatics and Robotics (IAR)
Artificial Intelligence for Language Technologies Lab (AI4LT)

Reviewer:	Prof. Dr. Jan Niehues
Second reviewer:	Prof. Dr. Alexander Waibel
Advisor:	M.Sc. Sai Koneru

01. February 2023 – 01. June 2023

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

. . . . . .

(Tobias Kaiser)

Abstract

This study aims to investigate methods for inducing document-level context awareness in Neural Machine Translation (NMT) models for bilingual chat data. Previous research suggests that contextual information is especially useful in the chat-domain [9]. Thus, we used Tiedemann's concatenation approach [41] to conduct two central experiments with different combinations of context.

In the first experiment, source context is prepended to the sentence that gets translated. Targeted Masking using an external language model is then applied to the source sentence, forcing the model to use context to resolve the masked information. We used a pretrained `M2M100` model and implemented two additional training stages, first training with large-scale, non-conversational data and then with a smaller, chat-specific dataset. Models were trained and evaluated on English-German and English-Chinese data.

The results show that in normal inference, no document-level model can score higher than the sentence-level baseline regarding BLEU. But the document-level model only trained with first-stage masked data obtains a significant result in contrastive evaluation using the ContraPro dataset. It outscores comparable Transformer-based models from the original ContraPro work by 8 percentage points.

The second experiment exploits context from the other speaker of the conversation, to predict the correct German formality level of a conversation. Therefore, a new dataset is proposed, by changing the German formality level of around half of the conversations in the existing BConTrasT chat dataset. Our results illustrate, that the model trained with target context outscores the sentence-level baseline concerning BLEU (+3.5) and the correct choice of formality level (+28 percentage points).

Zusammenfassung

In dieser Thesis werden Methoden zur Induktion von Kontextbewusstsein in maschinellen Übersetzungsmodellen für zweisprachige Chat-Datensätze untersucht. Vorherige Forschung haben gezeigt, dass kontextbezogene Informationen besonders bei Chat-Konversationen nützlich sind. Daher führen wir zwei zentrale Experimente mit verschiedenen Kombinationen von Kontext durch. Wir nutzen Tiedemann's einfachen Ansatz, um Kontext in den Übersetzungsprozess miteinzubeziehen.

Im ersten Experiment werden vorherige Sätze desselben Sprechers mit dem aktuell zu übersetzenden Satz verkettet. Anschließend wird der zu übersetzende Satz mithilfe eines externen Sprachmodells gezielt maskiert. Das Modell wird dadurch gezwungen, den Kontext zu verwenden, um den die maskierten Wörter übersetzen zu können. Als Modell haben wir ein vortrainiertes `M2M100`-System genutzt und zwei zusätzliche Trainingsstufen implementiert. In der ersten wurde das Modell mit einem großen parallelen Datensatz trainiert, der allerdings keine Chat-spezifischen Unterhaltungen enthält. Die zweite Stufe wurde dann für das Finetuning mittels Chat-Daten genutzt. Die Modelle wurden mit Englisch-Deutschen und Englisch-Chinesischen Daten trainiert und evaluiert.

Unsere Ergebnisse zeigen, dass bei Inferenz mit normalen, unmaskierten Daten kein Modell, welches mit Kontext trainiert wurde, besser abschneidet als das Modell auf Satzebene. Allerdings erreicht das Modell, das nur mit maskierten Daten der ersten Trainingsstufe trainiert wurde, ein signifikantes Ergebnis bei der Auswertung mit einem kontrastiven Test-Datensatz. Es übertrifft vergleichbare Modelle aus der original Arbeit, die zum Test-Datensatz gehört, um 8 Prozentpunkte.

Das zweite Experiment nutzt den Kontext des anderen Sprechers der Konversation, um die korrekte deutsche Formalitätsstufe der Unterhaltung zu bestimmen. Hierfür wird ein neuer Datensatz gebildet. Dieser basiert auf dem existierenden BConTrasT Datensatz und modifiziert die deutsche Formalitätsstufe von etwa der Hälfte der Dialoge im ursprünglichen Datensatz. Das mit Kontext des anderen Sprechers trainierte Modell erzielt deutlich bessere Ergebnisse als ein normales, satzbasiertes Modell. Es erreicht +3.5 BLEU Punkte mehr und bei der korrekten Wahl der Formalitätsstufe erreicht es +28 Prozentpunkte Genauigkeit.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Multilingual Communication plays a pivotal role in today's globalized world. Translating speech or text has therefore become a fundamental task in numerous areas of society. Since manual translation done by humans is rarely available, costly and comparatively slow, methods to automatically do the job have been subject to a whole independent field in the research area of Natural Language Processing (NLP). So called Machine Translation (MT) proposes systems that are able to automatically convert an input sequence into its corresponding translation using computers in a short and efficient way. Ideally these generated translations should match the quality of translations by humans.

First works were using rule-based approaches where manually created rules resolved ambiguities and translated the sentence structure. These systems were soon to be found impractical as they do not scale well in more complex settings. Corpus-based MT techniques emerged as the next step. They comprise statistical MT (SMT) as well as neural MT (NMT) and use machine learning methods to learn translation directly from training data.

NMT systems have gained popularity in this field in recent years since performance has improved significantly. Feed-forward neural networks first started to get deployed in traditional statistical machine translation systems to re-rank possible translation results in the target language.[33] Models that also took the source sentence into account followed soon after [32]. Ultimately, systems using only a single neural network that directly transforms the source sentence into the target sentence achieved better performance than statistical models, as they overcome several drawbacks of SMT (e.g. curse-of-dimensionality, 0-probabilities).

Despite the outstanding progress in NMT recently ([42]), especially downstream tasks, like translation of conversational data, still are issues for modern systems. Generating translations with the correct choice of pronouns, lexical consistency and general coherence can be challenging and is particularly relevant in the task of bilingual chat translation. Corresponding data normally consists of short, noisy messages referencing each other over several turns. Typical phenomena occurring in discourse-based text comprise anaphoric pronouns, word ambiguity and reference chains.

To be able to resolve ambiguities and obtain a fluent translation of high quality, it has been pointed out that the context before the current utterance contains important information regarding the translation of the current sentence. NMT systems thus must base

their current prediction not only on the current input sequence, but also on previously spoken sentences. This is called document-level (DOC) translation, in contradiction to sentence-level (SEN) models that are only taking the present source sentence into account. Research in how contextual information can be effectively learned to improve translation results has become more and more important.

1.2. Goal of this work

Despite document-level context being particulary relevant in a chat-translation environment, most NMT systems still operate on the sentence-level. They simply translate sentence by sentence without considering previous source and target utterances. This work's goal is to propose methods that teach NMT models a deep comprehension of the document-level context and the coherence between succeeding sentences.

Since we work with conversational data, numerous possible variations of contextual information can be used. Throughout this research, fine-tuning of pretrained NMT models is conducted using context from the same speaker (source)and the other speaker (target), combined with targeted masking of the current source sentence. It is important to note that these methods focus on the general objectives during training and do not imply any structural changes of the model architecture.

2. Fundamentals

In this section, fundamental concepts and methods are described that are essential for understanding the following experiments. First, the general structure and functionality of Artificial Neural Networks (ANN) are explained. Next, the specific Encoder-Decoder and Transformer architectures are elaborated. Finally, typical data preprocessing and evaluation methods are presented.

2.1. Artificial Neural Networks

Artificial Neural Networks (ANN) can be seen as a nonlinear function which transforms a set of input values to a set of output values. The specific mathematical function is determined by a set of parameters, also called weights. The role of the weights can be best elaborated through looking at the smallest computational unit of an ANN, the perceptron.

2.1.1. Perceptron

A perceptron takes input values $X = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, a constant $w_0 \in \mathbb{R}$ and has a set of weights $W = (w_1, w_2, ..., w_n) \in \mathbb{R}^n$ with the same size as the input vector X. The transformation from input X to output o is conducted by taking the weighted sum v of all inputs x_i and applying it to an activation function σ:

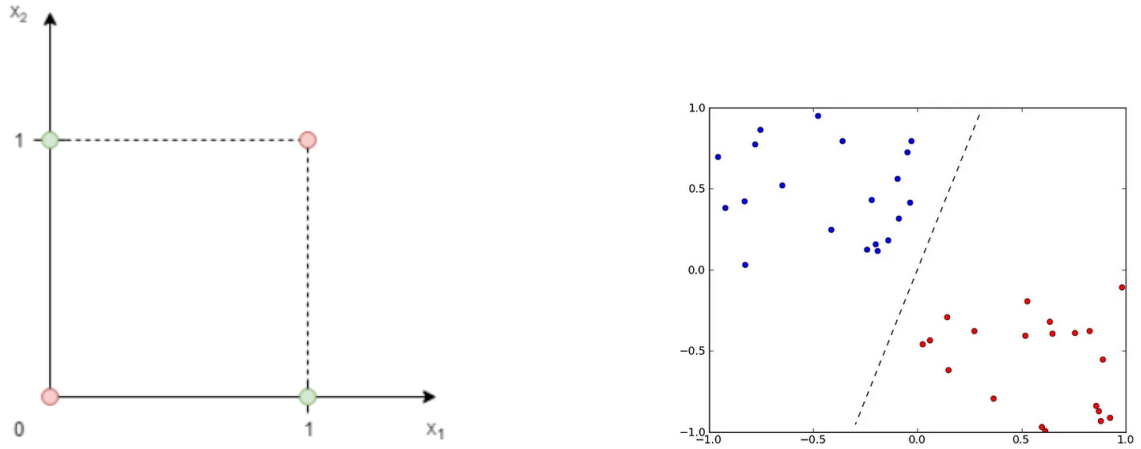
$$v = w_0 + \sum_{i=1}^{n} w_i x_i$$

$$o = \sigma(v)$$

Here, σ is typically a non-linear function that maps the result of v to a value between $(0, 1)$ or $(-1, 1)$, depending on the specific activation function that was chosen. The constant w_0, also called 'Bias', allows us to shift the activation curve up or down. The perceptron is often used as a binary linear classifier. A classifier maps the input vector to a predefined class as shown in Figure 2.1b. The perceptron therefore defines a $(n-1)$-dimensional plane that separates two classes. This classification comprises several drawbacks, which become evident in the case of the XOR-Problem. It is not possible to separate the two classes correctly using only one line.

2.1.2. Multi-Layer Perceptron

ANNs are nothing more than multiple perceptrons connected in series in multiple layers, so the output of one layer is the input for the next one. An exemplary structure of this



(a) XOR-Problem (b) Binary Classifier

Figure 2.1.: **XOR-Classification Problem**: It is impossible for a perceptron to classify each point correctly.
Binary Classifier: Input vectors left of the separating line are mapped to the "blue" class, vectors on the right are mapped to the "red" class (Image from [35])

so called Multi-Layer Perceptron (MLP) is depicted in Figure 2.2. Weights connect the neurons of consecutive layers, and what was the single output value of a perceptron now becomes a multidimensional output vector. Classification is now no longer limited to one separation plane.

The flow of information between two layers is structurally similar to the perceptron algorithm: We define the set of weights between two consecutive layers L_{i-1}, L_i as a $l_i \times l_{i-1}$-Matrix $W_{i-1,i}$. Having information present at layer L_{i-1}, the output, and thus the activation with activation function σ of layer L_i is calculated as follows.

$$L_i = \sigma(W_{i-1,i}L_{i-1})$$

The output of the whole transformation (with n Layers) can be described recursively:

$$L_1 = \sigma_1(W_{0,1}x)$$
$$L_i = \sigma_i(W_{i-1,i}L_{i-1})$$
$$O = L_n = \sigma_n(W_{n-1,n}L_{n-1})$$

As multiple separation planes are now possible, a MLP can conduct non-linear classification of data and therefore solve the XOR-Problem.

2.1.3. Training of an MLP

The non-linear function performed by a MLP is determined by the weights between its layers. All weights are typically initialized randomly. The best possible values for these

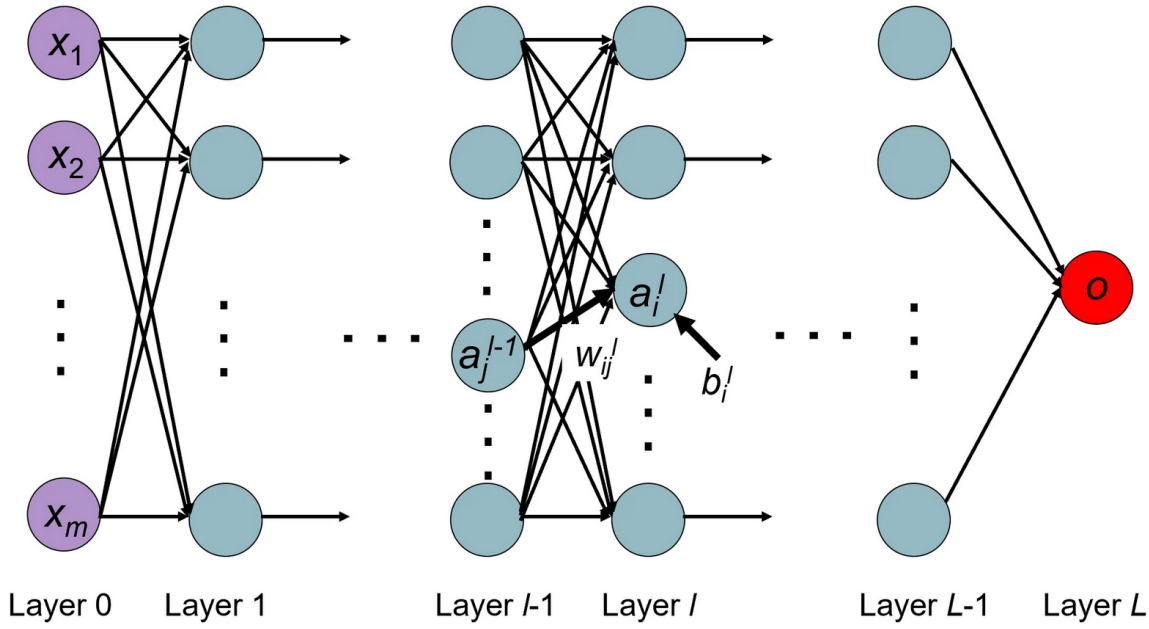


Figure 2.2.: MLP structure with input layer L_0, consecutive hidden layers L_{i-1}, L_i and output layer

weights are computed during the training process. This process is also known as fitting the MLP to the data, which is done by different learning algorithms. Learning algorithms are characterized by the usage of the prediction that is compared to the target output and by the adaption of parameters as a result to the former comparison [13].

Supervised Learning requires labeled training data, meaning target output is given for each input x. In an iterative process, an error function is used to compare the network output o to the target t for every input datum. Weights are then changed based on the measured difference between o and t according to the rule of a defined learning algorithm, such as *Stochastic Gradient Descent*.

Stochastic Gradient Descent aims for minimizing the error function E. In order to find a local minimum, the gradient $\frac{\delta E}{\delta w}$ with respect to the weights w is calculated. Each weight w_i is then modified in the opposite direction of its partial derivative $\frac{\delta E}{\delta w_i}$:

$$w_i^{t+1} = w_i^t - \gamma \frac{\delta E^t}{\delta w^t}$$

Here, γ stands for the learning rate, a parameter that is set before training determining the magnitude of the weight change. This process of backtracking the influence of each weight on the error and modifying them accordingly is called backpropagation algorithm. It finally stops if a pre-specified criterion is fulfilled, e.g. if the error is not decreasing or the difference between consecutive error values is below a certain threshold.

Unsupervised Learning works without given target labels. The task is to find pat-

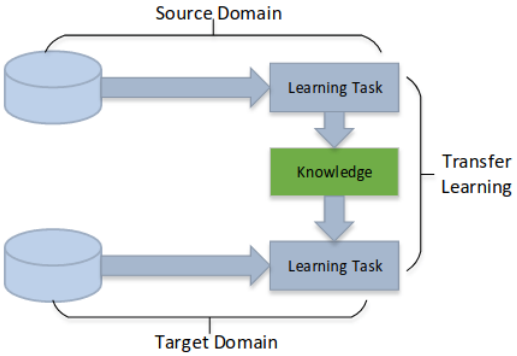


Figure 2.3.: Transfer Learning process between two domains [39].

terns in the dataset that could indicate differences between data-clusters. These patterns then are useful in further analysis and interpretation of the data.

2.1.4. Transfer Learning

Transfer learning is an important method for solving the problem of scarce training data in machine learning. In contradiction to traditional machine learning, knowledge from previously learned tasks is leveraged for training new models on more specific, downstream tasks with significantly less training data available (Figure 2.3). In the case of this paper, models that are pre-trained on basic bilingual machine translation data are used as a starting point. They are further trained, also known as fine-tuned, on the downstream task of bilingual chat translation.

Models trained using a transfer learning approach were found to perform significantly better in situations with insufficient data than models trained traditionally, isolated on the downstream task. Additionally, training time in the target domain is reduced [39].

2.2. Sequence-to-Sequence Models

Sequence-to-sequence (Seq2Seq) models are ANNs that take a variable-length sequence of items (e.g. letters, audio signals, words) as input and transform it to another variable-length sequence of items. This fits various use cases. Seq2Seq models are producing state-of-the-art (SOTA) results in numerous NLP tasks like Neural Machine Translation (NMT), Summarization [38], Speech Recognition [4] and many more. In the particular case of NMT, the input is a sequence of words and the output is the translated input sequence. Seq2Seq models consist of an encoder and decoder. The corresponding structure is explained in the following paragraphs.

Encoder-Decoder Structure Encoder-Decoder networks first got introduced by [37]. They tackle the tasks of input representation and output generation separately. The encoder is responsible for reading the input sequence and creating a memory vector where all

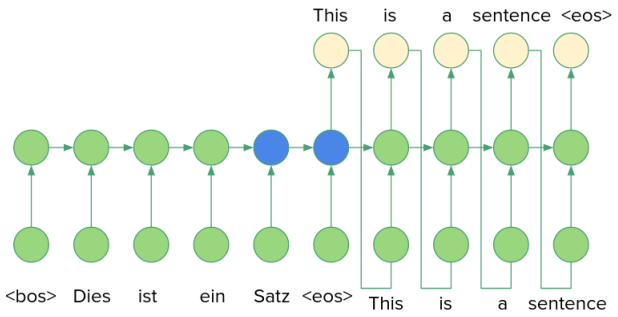


Figure 2.4.: Encoder-Decoder architecture: The encoder reads the input sequence $X = (x_1, x_2, ..., x_m$ and creates a state representation $Z = (z_1, z_2, ..., z_m)$. Memory vector $z_m$is passed to decoder (light-blue). The decoder generates the output sequence $Y = (y_1, y_2, ..., y_n)$ word by word. To predict output y_{i+1} the decoder calculates a new hidden state h_{i+1} based on the last state h_i as well as z_m and an embedding of the previously predicted word y_i [12].

contents of the input sequence are stored. The decoder then decodes this memory vector and generates the output sequence autoregressive word by word. The current output thus doesn't solely depend on the previously generated words, but also on the input sequence read by the encoder (Figure 2.4).

A potential drawback of the basic Encoder-Decoder architecture is that all information of the input sequence is squashed into one fixed-size memory vector. This leads to a potential loss of valuable information and makes it hard for the network to cope with long sentences [1]. Previous studies showed that performance plummets when working with sentences that are longer than the sentences in the training corpus [5].

In order to address this issue, [1] proposed a technique of aligning source and target sequences during translation. The so called "Soft-Search" or "Attention" mechanisms are further explained in the next section.

2.3. Transformer

Many of the previously discussed drawbacks of a basic Seq2Seq system are remedied by the so-called Transformer architecture. The Transformer is a model that relies completely on Attention mechanisms, making the recurrent structure of a Seq2Seq model obsolete. Functionality of Attention and the original Transformer architecture are explained below.

2.3.1. Attention

Attention mechanisms were first combined with Long Short-term Memory (LSTM) Networks by [1]. LSTM nets can be used in Seq2Seq architectures and are a type of recurrent neural network (RNN) that are specifically designed to handle long-term dependencies in sequential data. For every generated word, proposed model from [1] performs a search on

the input sequence, looking for positions where relevant information is concentrated. It then uses the obtained context vector associated with the source positions and predicts the next word based on it, as well as the already generated target sequence. In contradiction to the vanilla encoder-decoder approach, the model does not try to encode an input sequence into a single fixed-size vector. It rather converts it into a sequence of vectors from which a subset is chosen for generation of the next word.

[1] reported results that show better translation performance compared to basic encoder decoder models, especially regarding the translation of long sequences. They show that the general approach to focus only on selected parts of the input benefits translation quality of long-term dependencies. However, some pitfalls remain. Conducting the search for every generated word can be computationally expensive, especially for longer sequences. In addition, the recurrent architecture is difficult to parallelize and thus slowing down training and inference significantly.

By introducing the **Transformer** architecture, [42] solved these issues. For their proposed model, they combine different attention mechanisms. Similar to [1] they allow the decoder to focus primarily on relevant parts of the source sentence while making a prediction. Therefore, the whole encoder state representation Z is being kept to compute a context vector c for prediction in the decoder [12].

$$c_i = \sum_{j=1}^{m} \alpha_i j z_i$$

Attention score α is the result of an *alignment model*. This neural network model is trained jointly with the Seq2Seq model, that measures how well input i is aligned with the input at time step j. Scores for every input form an Attention matrix, showing intuitive relations between input and prediction (Figure 2.5).

Multi-Head-Attention performs multiple attention functions in parallel, using multiple attention heads. Resulting context vectors are concatenated and linearly projected. This approach was found beneficial as it allows the model to jointly attend to information from different representation subsets at different positions. Averaging over a single attention head prohibits this [42]. In contrary to calculating attention scores between two distinct sequences, **Self-Attention** works on just one. Relating different positions of the same sentence, the goal is to obtain an attention-based representation of the sequence [42].

2.3.2. Transformer Architecture

The Transformer architecture combines an Encoder-Decoder architecture with stacked self-attention, multi-head attention and fully connected layers depicted in Figure 2.6. Encoder and Decoder are composed of N identical layers (in the original paper [42] $N = 6$). One en-

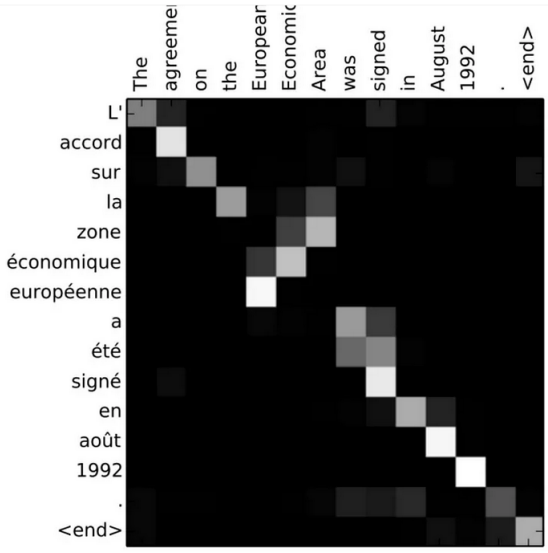


Figure 2.5.: Attention matrix: Shows alignment between French source sentence and English translation. Each pixel shows alignment scores α_{ij} for the i-th input and the j-th output word (black: 0 to white: 1). This graphic depicts the different order of object and adjectives in both languages very well. (Image from [7])

coder layer comprises two sub-layers. The first one implements a multi-head self attention function and the second is a simple fully connected feed-forward neural network. Similarly, a single decoder layer contains a self-attention and a feed-forward part, whereas masking is applied at the self-attention level to prevent attention to subsequent positions. Additionally, a third sub-layer is added to perform multi-head attention over encoders output. Each sub-layer of both encoder and decoder is surrounded by residual connections, followed by layer normalization. Thus, $LayerNorm(x+Sublayer(x))$ is the output of each sub-layer.

Transformer models can be trained significantly faster than structures based on recurrent layers and achieve outstanding results in translation tasks. They overcome the problem of learning dependencies between distant positions by utilizing attention to reduce the number of operations between two arbitrary input and output positions to $O(1)$.

2.4. Data Preprocessing

Before raw data like text or images can be used as input for an ANN, several pre-processing steps have to be applied. In the following, the basic pipeline from raw sentences to binary input for our ANN models are elaborated.

2.4.1. Tokenizing

Tokenization describes the process of dividing text into smaller, more basic units, called tokens. Without these basic units clearly separated, the input can not be interpreted

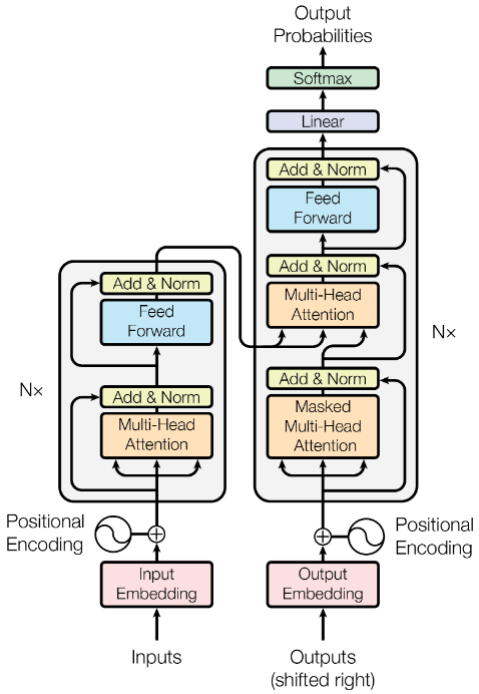


Figure 2.6.: Transformer architecture [42]

correctly during training phase. Identification of tokens depends on the language and the tokenization technique used. In English, simple white space delimiters can be utilized to split a sentence into words (White Space Tokenization), whereas in Chinese, the absence of such delimiters enforce other tokenization approaches [44].

Subword tokenization is splitting text into even smaller units, splitting less frequent words into subwords. Two of these methods are explained in further detail since they are used before training different models in this work.

2.4.2. Byte-Pair Encoding

Byte-Pair Encoding (BPE) was introduced by [11] and initially served as text-compression algorithm [36]. It was first used in a NMT context by [34].

The general functionality of BPE is to ensure that the most common words in the data are represented as a single token in the resulting vocabulary, whereas rather rare words are split into two or more subword tokens. The algorithm therefore first splits each word into characters and counts their occurence. After that, the most frequent character-pairing is searched and merged together, creating a single token out of two tokens. This is done iteratively until a given token or iteration limit is reached. Merging finally leads to a corpus with the least number of tokens [17].

2.4.3. SentencePiece

SentencePiece tokenization uses the BPE algorithm as a basic segmentation method, but provides additional features that benefit the overall tokenization process. The SentencePiece algorithm implements tokenization on a raw input stream of unicode characters rather than on white-space-separated words, treating spaces as normal characters. Consequently, there is no language-dependent logic and tokenizing languages without explicit space delimiters (e.g. Chinese or Japanese) is possible without additional pre-tokenization steps. Additionally, the algorithm works with a predetermined number of unique tokens, resulting in a fixed vocabulary size of e.g. 8k, 16k or 32k. This makes it very applicable for typical NMT models that operate with fixed vocabulary.

Another distinctive aspect of SentencePiece is the implementation of subword regularization. As subword segmentation is potentially ambiguous, multiple word separations are possible even with the same vocabulary. It was found that segmentation ambiguity could be used as a noise to improve NMT robustness and accuracy. Therefore, multiple subword segmentations are sampled probabilistically during training [20].

2.5. Evaluation Methods

Performance of NMT models needs to be properly measurable and comparable between different models. Human evaluation is both time-consuming and expensive. Numerous automatic evaluation methods were proposed in the past, using individual techniques for comparing model prediction with output target. The metrics we used during our research are discussed below.

2.5.1. BLEU

The Bilingual Evaluation Understudy (BLEU) was proposed by [28] as a quick, inexpensive and language-independent NMT evaluation method that correlates highly with human evaluation. BLEU score indicates the "closeness" of a translation generated by a NMT model to a professional human translation.

The primary function of BLEU is the comparison of n-grams of predictions with reference translations and count the number of matches. The resulting precision scores p_n (see Figure 2.7) for each n-gram are combined to the Geometric Average Precision (GAP) and penalized with a brevity penalty (BP), to punish short predictions [6].

$$\begin{aligned} GAP(N) &= \prod_{n=1}^{N} p_n^{w_n} \\ &= (p_1)^{w_1}(p_2)^{w_2}\cdots(p_N)^{w_N} \end{aligned}$$

$$BP = \begin{cases} 1, c > r \\ e^{\frac{1-r}{c}}, c <= r \end{cases}$$

Figure 2.7.: Illustration of 4-gram comparison. Here $p_4 = \frac{2}{5}$

$$BLEU(N) = BP \cdot GAP(N)$$

with

- N = maximal N-gram a precision score is calculated for
- $w_n = \frac{1}{N}$, uniform weights
- c = length of predicted sentence
- r = length of target sentence

Drawbacks of BLEU comprise missing consideration of the word meaning and ignoring the importance of words. Despite that, it has become one of the most widely used evaluation metrics, making it easier to compare results with other work.

2.5.2. COMET

COMET (Crosslingual Optimized Metric for Evaluation of Translation) was proposed by [30]. It serves as a framework for training individual NMT evaluation models, but also provides pretrained models implementing different kinds of metrics. Additional to the hypothesis of an NMT model and the reference sentence, COMET estimator models also take the source sentence into account during evaluation.

Using a neural network for evaluation purposes alleviates the problem of exact word matches, on which metrics like BLEU rely on. The basic architecture of an estimator model is depicted in Figure 2.8. [30] use a pretrained XLM-RoBERTa model for the encoder and train a system on top of that for scoring. Source sequence, hypothesis and reference are fed into the pretrained cross-lingual encoder. Resulting word embeddings are then passed through a pooling layer, generating sentence embeddings for each of the three input sequences. The following concatenation layer combines and concatenates the sentence embeddings into one single vector that is finally fed to a feed-forward regressor. Basic Mean Squared Error (MSE) is calculated and minimized to train the model.

When training a COMET model, a target evaluation score is necessary to compare model output to. Thus, by striving for a specific metric, models can be trained to implement different types of human judgements. In the original work, three exemplary models were trained on scores generated by *Direct Assessments, Human-mediated Translation Edit Rate* and *Multidimensional Quality Metrics* [30].

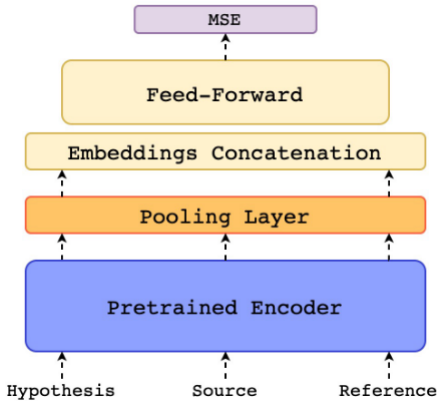


Figure 2.8.: Estimator model architecture, graphic from [30]

3. Document-level Pretrained Chat-MT

This section showcases the different approaches to include document-level (DOC) context in Neural Chat Translation (NCT) systems. First, the used datasets are presented. After that, the `M2M100` model is introduced as the central model on which the following training methods are implemented. Next, we explain how we incorporated context in the training process. Finally, the two major experiments of this study are elaborated: Targeted masking and exploiting target context.

3.1. Datasets

In the following, more information about the datasets used to train NMT models in this work is given. All of them are so-called parallel, bilingual datasets. In NMT, parallel data refers to a collection of texts that are aligned with each other. This means each sentence has a corresponding translation in another language.

3.1.1. News-Commentary

One very popular parallel dataset is the News-Commentary corpus from OPUS [40]. It comprises 15 languages and 109 parallel bitexts. The set consists of news articles and their corresponding commentaries, covering a wide range of topics, including politics, economics, science, and technology, among others. Since experiments in this work are conducted on language pairs {en, de} and {en, zh}, the corresponding stats can be found in Table 3.1. We used the dataset version provided by the Huggingface Hub.

News-related texts contain various discourse phenomena that are common in natural language (e.g. anaphoric pronouns, ambiguous words or other types of referring expressions). Hence, this corpus was used in several papers studying translation of conversational data (e.g. in [14]). Anaphoric pronouns, in particular, are frequently used to refer back to previously mentioned entities or concepts, making them an important phenomenon in the course of this work.

3.1.2. BConTrasT

Moving to conversational datasets, more specifically chat conversations, BConTrasT is used to finetune the models in this work on the downstream task. The corpus was introduced in the WMT 2020 shared task on chat translation [9]. It is based on the monolingual English *Taskmaster-1* set [3] and includes task-based dialogs in six different domains:

News-Commentary	en,de	en,zh
#Sentence Pairs	223K	69K

Table 3.1.: Size of news-commentary dataset

	en-de		de-en		en-zh		zh-en	
	lines	dialogs	lines	dialogs	lines	dialogs	lines	dialogs
Training	6.216	550	7.629	550	5.560	1.036	4.427	1.036
Dev	862	78	1.040	78	567	108	517	108
Test	967	78	1.133	78	1.466	274	1.135	274

Table 3.2.: Size of chat datasets

1. ordering pizza
2. setting appointments at an auto repair garage
3. booking an Uber
4. buying movie tickets
5. ordering coffee
6. making restaurant reservations

A subset of the monolingual corpus was translated into German using the Unbabel Translation Service. To produce data with explicit discourse phenomena, the selected conversations contain the English anaphoric pronoun *it* at least once[9]. Since all dialogs concern sevices between two parties, speakers are either *customer* or *agent*. In BConTrasT, the customer is always the German speaker, requesting a certain service. The responding agent speaks English. Size and statistics can be found in Table 3.2.

3.1.3. Formality-BConTrasT

This work additionally proposes a new dataset obtained by switching formality level of pronouns in about half of the dialogs in BConTrasT. The **Formality level** of a German sentence refers to the two possible translations for the English pronoun *you*. For example, the sentence *And what time would you like your reservation?* can be translated to one of the succeeding German translations:

Und für welche Zeit möchten Sie Ihre Reservierung? (2)

Und für welche Zeit möchtest du deine Reservierung? (3)

Translation (2) depicts the polite form, $\{you \rightarrow Sie, your \rightarrow Ihre\}$ whereas (3) would be used if both speakers are on a first-name basis $\{you \rightarrow du, your \rightarrow deine\}$. In the next

Können Sie bitte bestätigen, dass Sie diese Adresse haben?

↓

Kannst du bitte bestätigen, dass du diese Adresse haben?

Figure 3.1.: Original *Sie-level*-sentence and modified sentence. Marked are the 3 verbs, the first two get modified correctly whereas the third one is falsely conjugated (right form would be *hast*).

sections, sentences that are on the same formality level as (2) will be referred to as *"Sie-level"*-sentences, contrary to sentences like (3) which will be called *"Du-level"*-sentences. Sentences without the use of the pronoun *you* are regarded as neutral and are insignificant for the production of Formality-BConTrasT.

Originally, the BConTrasT corpus comprises conversations on *Sie-level* only. Since the choice of formality level normally is agreed on in the beginning between the two involved speakers, randomly switching half of the dialogues pronouns provides a more realistic setting.

Basic steps of the dataset-transformation include iterations over all conversations, each of which has a 50% chance of getting selected for level-switch. As the switch exclusively affects German sentences, only source sentences of the customer as well as target sentences of the agent need to be modified. *Sie-level*-sentences in selected conversations are altered by a basic, greedy algorithm using regex to substitute matching patterns. Hence, grammatical mistakes occur occasionally regarding the correct conjugation of verbs (Figure 3.1). This is considered ignorable for evaluation on pronoun precision, but could affect automatic metrics like BLEU and COMET in small manners.

3.1.4. BMELD

BMELD (bilingual MELD) [22] is a bilingual English-Chinese chat translation corpus based on the MELD (Multimodal EmotionLines Dataset) dataset. The original monolingual MELD corpus is a combination of the EmotionLines dataset, which consists of dialogues from movies annotated with emotional labels, and corresponding audio data from the TV show *Friends*. It has been used in a wide range of research on multimodal sentiment analysis and emotion recognition.

Monolingual English data was automatically translated and manually post-edited to obtain a parallel bilingual corpus. BMELD provides conversations between two or more speakers. We want to simulate only two speakers, like the setting in BConTrasT. Thus, we divided the several different speakers into two groups with around the same number of utterances per group. One group of speakers then got assigned the "customer" language direction (Chinese→English) and the other group the "agent" language direction. Additional meta-information like sentiment or emotion of the utterance are existent but ignorable. That's because this study conducts no experiments concerning sentiment analysis or bilingual conversational characteristics. Stats can be found in Table 3.2.

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

Table 3.3.: Contrastive Evaluation: *It* refers to a bat or *Fledermaus (f.)* which makes *sie* the correct reference translation. Example taken from [26]

3.1.5. ContraPro

ContraPro is a large-scale test set with focus on specific discourse phenomena like pronoun translation. More specifically, it automatically evaluates the accuracy with which NMT models translate the English pronoun *it* to its German counterparts *er, sie* and *es* [26]. This ambiguity can only be resolved using contextual information, making ContraPro a well-suited corpus for testing the context-awareness of models.

ContraPro provides data for **contrastive evaluation**. For each English source sentence, three German target translations are provided, each with a different translation of *it*. The correct translation is defined by the antecedent, the object which the pronoun refers to. Antecedents occur in the context sentences provided by the dataset and are categorized by their distance to the source sentence. A distance of 1 means that the antecedent is part of the directly preceding context sentence.

To obtain a pronoun accuracy, each reference and contrastive translation needs to be scored and compared. Since NMT models are in fact language models of the target language conditioned on the source input, they can be used to calculate a probability score for an existing translation. If the model assigns a higher probability score to the actual reference translation than to the contrastive examples, this case is referred to as a "correct decision" by the model. It is important to note that this refers only to the right pronoun choice. When letting the model generate a new translation given the source sentence, it may produce a completely different target sequence compared to the reference.

Another possible evaluation method is generative evaluation (e.g. used in [29]). Here, the words before the pronoun are given, and it is checked if the model generates the correct pronoun. This is not part of the metrics used in this work.

3.2. Baselines

This section describes the `M2M100` as it serves as the base-model for various fine-tuning approaches in this thesis. Furthermore, models evaluated in the beginning of this research are shortly introduced.

3.2.1. General Baselines

In the beginning of the research, several popular Seq2Seq models have been evaluated on chat data to have a greater set of comparable systems. These models have not been fine-tuned on the news-commentary, BConTrasT or BMELD dataset. Instead, a pretrained checkpoint available from Huggingface was used during inference. Next are short descriptions of the different models.

MBART (Mulitlingual Bidirectional and Auto-Regressive Transformers) was presented in [23] and is a multilingual Seq2Seq model primarily trained for the translation task. It was trained on large-scale monolingual data in various languages using the BART objective. The original BART model used a denoising auto-encoding objective, meaning the model was trained to reconstruct the original input sequence from a corrupted version.

FSMTwmt19 (FairSeq Machine Translation) is Facebook's submission to the 2019 WMT News-Translation task [27]. It is a transformer-based architecture trained with the Fairseq sequence modeling toolkit. The model achieved state-of-the-art performance on multiple language pairs, including English-German.

Raw-Transformer is a NMT model built purely like the architecture described in the original "Attention is all you need" paper [42], using an annotated version from [18]. It was not pretrained in any manner.

3.2.2. M2M-100

M2M100 is a multilingual Transformer-based model trained for Many-to-Many (M2M) multilingual translation. It was proposed by [8] with the aim to overcome drawbacks of typical english-centric NMT-models by conducting translation directly between 9,900 directions of 100 languages. M2M-100 is not specifically trained for conversational data but achieves good scores with automatic metrics. It was also used as a baseline for comparison in the WMT 2022 shared task on chat translation [10]. It is pretrained on the language pairs used in this work and functions as a starting point for the models fine-tuned on downstream data. Note that [8] provide three different model-"sizes" regarding the number of parameters. The following studies are all based on the small model with 418 million parameters.

Since utilization of context is the central point of this study, two separate baselines were trained:

- *Sentence-level* (SEN): Each sentence is passed separately to the model
- *Document-level* (DOC): Along with the sentence s that gets translated, N context sentences are passed to the model. Context either fully consists of the last N utterances of the same direction as s (source-context) or the opposite direction (target-context). Only s gets translated by the model.

Solely training on scarce chat-data (BConTrasT/BMELD) was not sufficient for learning to use information out of the context. Thus, an additional training stage with a larger but non-conversational dataset (news-commentary) was implemented to enable DOC-level learning.

3.3. Concatenation Approach

To achieve the aim of document-level understanding, we first must be able to feed previous sequences to an NMT model. There are numerous methods to do that. A few of them are mentioned in section 5.2. This section describes a simple concatenation approach which is later used during training and inference of the `M2M100` model.

The simplest way to incorporate contextual information in the current translation is to prepend context from the previous utterances to the sentence to be translated. It was, among others, done by [41]. This approach does not change the model structure, and implementation is easy and straightforward. Concatenated context can comprise either previous sentences from the same speaker (source-context, same language) or from the other speaker (target-context, different language) only.

Practical implementation is depicted in Figure 3.2. Sentences are separated by a special token *SEP*. To avoid ambiguities, *SEP* should be used exclusively for separation and should not occur in any of the languages of the sentences. Additionally, *SEP* should already have an embedding in the pretrained system. This is ensured by choosing a single Korean character out of the vocabulary used during pretraining. Thus, otherwise needed enlargement of the embedding layer is prevented.

Prior to any preprocessing step, data is provided in raw text files with one utterance per line. At this state, there is no syntactical way to distinguish end and beginning of consecutive dialogs. But clear boundaries are needed for selecting only sensible context when concatenating. E.g. the first sentence of a conversation does not have any prior available contextual information, so no utterances should be prepended during the concatenation process. An *EOD* token (End-Of-Dialog) inserted at the end of each conversation provides is used to make data distinguishable. As with the *SEP* token, *EOD* is a single symbol out of the pretraining vocabulary.

We hypothesize that in the domain of chat messages, relevant context mostly occurs in the utterances directly before the current source sentence. In order to limit computational cost, context size should be chosen as efficiently as possible. Hence, in the experimental setup of this work, two to four context units are used. Models trained with corresponding context size are marked with the keyword **3to1**, referring to the number of sentences on the source and target side.

Hi, how are you? <*SEP*> Great, thanks for asking, how can I help you?
↓
Sehr gut, danke für die Nachfrage, wie kann ich Ihnen helfen?

Figure 3.2.: Source sentence with one context sentence. Gets translated to one target sentence (2to1). A seperation token divides the sentences on the source side.

3.4. Experiments

With the method to include context in NMT systems set, we can move forward to explain the two central experiments conducted in this work. The following section first describes the targeted masking approach, a masking algorithm based on probabilities assigned by a LM. Targeted masking is evaluated on English-German and English-Chinese data. Following, using target utterances as context is explained. This is done in the English→German direction only, using the Formality-BConTrasT dataset.

3.4.1. Targeted Masking

Masking is an efficient data-augmentation method to obtain a more generalized model. Previous work proposed masking of random tokens in the source sentence to force the model to recover the missing information when translating [21]. Targeted masking however implies that choice of tokens for masking should rather be selective than random. We examine a selection based on language models (LM) for training data of both stages. LMs are pretrained but not finetuned on our downstream data. For German and English data, fairseq's pretrained neural language models were used.

General methodology of the masking algorithm is explained in Fig 3.4. The idea is to mask words that a LM would assign a higher probability value to if context is provided, compared to without any context. The resulting augmented data contains masked tokens. They should be resolvable by a NMT model using information given in inter-sentential context.

Since the LM calculates scores on the token level, several points have to be ensured during postprocessing to obtain valid masking:

- Only "real" words were masked, excluding tokens like sentence punctuation, phone-numbers, times, etc.
- Masking was applied to full words, meaning if any sub-token of a word was chosen for masking, the whole word was masked.
- Masking was only implemented on sentences to be translated, not on context sentences.

	Scores without context			
	w_1	w_2	w_3	w_4
	0.71	0.34	0.54	0.12
	Scores with context			
C	w_1	w_2	w_3	w_4
	0.68	0.92	0.57	0.10

Table 3.4.: Targeted Masking: LM iterates the same data twice, the first time on SEN-, the second time on DOC-level. During each iteration, LM calculates the positional probablility scores p_i of each word w_i in a sentence $S = (w_1, w_2, ..., w_n)$. The result are two scores per word: p_i^{SEN} and p_i^{DOC}. Finally, masking is applied to word w_i if $p_i^{DOC} - p_i^{SEN} > T$ with T as a constant threshold.

Through the value of threshold T, the number of masked words can be influenced. Whilst also inspecting effects of more masking, the results presented in Section 4.2 are based on data with statistics shown in Table 3.5.

3.4.2. Using target context

One characteristic property of bilingual chat-conversations is the division of utterances between two or more speakers. References across different speakers and languages are not uncommon. Hence, context-aware models should not only be incorporating source side, but also target side context from other speakers. Similary to source context being able to disambiguate pronouns or other words, target sentences can contain important information concerning gender of pronouns or formality level.

Formality-BConTrasT provides sentences with different German pronoun translations concerning the formality-level of the conversation. It basically implements the fact that translation of the pronoun *you* should depend on the pronoun that is being used by the German speaker (*Sie* or *du*), making it perfect for showcasing the use of target context. Context is prepended to the source sentence by using the concatenation approach elaborated earlier. A **3to1** model considers the last two German target utterances before the current source sentence as context. It does not matter if sentences of the source speaker occur between the sentences of the target speaker.

Models To demonstrate the effectiveness of target-context awareness, the central model that is actually trained with target context on Formality-BConTrasT is compared to following other systems:

- SEN-Baseline: A baseline trained on sentence-level data during both training stages.
- Sie-Baseline: A baseline that simulates a model translating conversations to *Sie*-level only. Instead of training a whole new system, the model's translation of the test

en→de	news-commentary			BConTrasT		
	train	valid	test	train	valid	test
#masked	156K	33K	33K	6.5K	862	1K
#total	3.5M	748K	757K	73K	9.5K	10.5K

Table 3.5.: Number of masked words in both training stages for ($en \rightarrow de$) direction

de→en	news-commentary			BConTrasT		
	train	valid	test	train	valid	test
#masked	152K	33K	33K	4K	586	668
#total	3.5M	748K	759K	41K	5.5K	6K

Table 3.6.: Number of masked words in both training stages for ($de \rightarrow en$) direction

zh→en	news-commentary			BMELD		
	train	valid	test	train	valid	test
#masked	62K	3.5K	3.5K	3.5K	368	1K
#total	7M	390K	394K	80K	8.5K	22K

Table 3.7.: Number of masked words in both training stages for ($zh \rightarrow en$) direction

en→zh	news-commentary			BMELD		
	train	valid	test	train	valid	test
#masked	59K	3K	1.6K	3K	347	756
#total	3M	147K	79K	40K	4.5K	10.5K

Table 3.8.: Number of masked words in both training stages for ($en \rightarrow zh$) direction

Customer	Agent
Hallo, kann ich Sie zu Ihnen sagen?	
Wie kann ich Ihnen helfen?	How can I help you?
Hallo. Ich suche nach Kinokarten.	Hi. I am looking for movie tickets.
Okay welchen Filem wollten Sie?	Okay what film did you want?
Glass	Glass
Der ist sehr beliebt	That's really popular

Table 3.9.: Example conversation from Formality-BConTrasT with added introducing sentence. The corresponding translations are depicted in blue.

set is synthetically constructed. This is done by taking the output of SEN-Baseline and modifying the pronoun formality level using the same technique as during the construction of Formality-BConTrasT.

- Du-Baseline: A baseline that simulates a model translating conversations to *Du*-level only. Similar to Sie-Baseline, no new model was trained, but output was produced manually.
- Mixed-Baseline: This is a mixture of Sie-/Du-Baselines. It randomly selects 50% of dialogs to change to *Du*-level, whereas the other half is modified to be on *Sie*-level.
- Rule-Based-Model: Acts as an upper limit for pronoun accuracy by simulating perfect context-awareness. As with Sie- and Du-Baseline, it modifies SEN output. The system scans the German context for pronouns or conjugated verbs that indicate the level of formality and changes the prediction accordingly. Assuming the algorithm is perfectly reliable in detecting the correct formality level, the resulting accuracy is the highest achievable with the used context size.
- DOC-TGT-Model: System that is trained with target utterances as context in the second training stage using Formality-BConTrasT. The first training phase was conducted on news-commentary using two source context sentences. Context size was increased in the second phase (*5to1*) to overcome the little use of pronouns by the target speaker.

Challenges lie in the structure of the BConTrasT corpus. Due to the service-oriented nature of the dialogs, the German-speaking customer generally does not use a lot of pronouns in his utterances. Sentences are rather short, and often are questions for information about restaurants, movies, etc. The use of pronouns was synthetically increased by adding a sentence at the beginning of each conversation (Fig. 3.9). The added sentence simulates a typical German introduction, which is defining the formality level by asking the other speaker for permission to use the according pronouns. A bigger context window was also used (5to1 in second stage training), to maximize the source sentences with a pronoun in their context, that reveals the formality-level of the dialog.

Summary In this section, we elaborated the two central experiments of this study: Targeted masking and using target context to detect the right formality level. We therefore took a look at the Tiedemann approach [41] to include preceding utterances in the translation process. In addition, the datasets we used for training were introduced and our baselines as well as the `M2M100` model were discussed. The results of the experiments are now presented in the following chapter.

4. Evaluation and results

The following part of this paper moves on to present the results of the conducted experiments. Evaluation of the models trained with targeted masking and target context are being reported, among several other results that were obtained during the execution of this research.

4.1. General baselines

As a starting point of this research, several pretrained or untrained models were evaluated on the BConTrasT chat data. MBART and FSMTwmt19 models were pretrained models loaded and evaluated from Huggingface. MBART-50 is a MBART model extended to comprise pretraining of 50 languages (original MBART was only pretrained on 25). A score for the pretrained checkpoint of M2M100 before any fine-tuning was applied is also reported.

	BLEU	
Model	en-de	de-en
mBart50	36.57	47.8
FSMTwmt19	42.87	49.05
Raw-Transformer	36.22	36.05
m2m100-418	32.16	34.01

Table 4.1.: General Baselines that were evaluated on BConTrasT at the very beginning of this research

4.2. Targeted Masking

Next are the results of training with targeted masking. Since training consisted of two training stages, all possible combinations of masked and unmasked data were evaluated. Different models are depicted by different two-character-combinations. E.g. *NM* refers to the model that was first trained with normal (unmasked) news-commentary data in the first stage, and after that was trained with masked chat data in the second stage. The following scores were obtained by evaluating with the test split from the chat datasets (BConTrasT and BMELD). Inference was also split into inference with normal data and

inference with masked data. Models were trained with context (DOC) as well as without (SL). All DOC-level models were trained with 2 context sentences (3to1). Inference of these models was also conducted with 2 context sentences. *NN* models are the SL-/DOC-Baselines.

4.2.1. Evaluation on Chat Data

English-German: As can be seen from Table 4.2, for normal inference the SL-Baseline outperforms other models regarding BLEU scores. The DOC-Baseline scores about −0.7 BLEU points lower. It seems raw context without any masking is causing more noise than it benefits the translation. With masked training in one of the two stages, better performance shifts to DOC-level models with the DOC-*NM* surpassing the corresponding SEN-level model by around +1.3 BLEU. Surprisingly, DOC-*MM* achieves the best COMET score in normal inference.

Unsurprisingly, *MM* models perform best in masked inference. Regarding BLEU, DOC-*MM* gains +0.4 points compared to the same model without context. But concerning COMET, SEN-*MM* achieves the best score by +1.6 points. DOC-*NN* scores way lower than SEN-*NN*. It seems to be overburdened by preceding context and masking tokens it has never seen during training.

	Sentence-Level				Document-Level			
Inference	NN (BL)	NM	MN	MM	NN	NM	MN	MM
Normal (BLEU)	**53.7**	52.01	52.76	53.12	53.02	53.35	53.39	52.88
Normal (COMET)	90.22	90.06	90.26	90.33	89.72	90.03	90.82	**90.92**
Masked (BLEU)	43.57	49.01	47.66	49.8	30.56	49.51	47.71	**50.29**
Masked (COMET)	81.56	88.03	85.65	**88.04**	72.52	85.74	84.26	86.36

Table 4.2.: Results of targeted masking *en* → *de*

Regarding German→English direction (Table 4.3), results are similar to the other direction. The sentence-level baseline performs best in normal inference concerning BLEU and COMET scores. The document-level baseline achieves about −2 BLEU points less. Document-level models with masking apparent in at least one training stage score slightly higher than DOC-*NN*. DOC-*MM* obtains the best BLEU score in masked inference, significantly surpassing all sentence-level systems.

English-Chinese: Looking at the English→Chinese direction, no document-level model can perform better than its sentence-level counterpart. Surprisingly, the SEN-*NN* model also obtains the best BLEU score for masked inference. So inheriting context and masking during the training process harmed the general translation quality of the other systems significantly. Here, no sign of sensible usage of contextual information can be seen.

	Sentence-Level				Document-Level			
Inference	NN (BL)	NM	MN	MM	NN	NM	MN	MM
Normal (BLEU)	**56.93**	52.43	56.17	51.84	54.72	55.33	54.88	55.89
Normal (COMET)	**92.42**	91.9	92.12	91.8	92.06	92.06	92.11	91.65
Masked (BLEU)	43.45	46.94	48.68	44.41	24.1	50.0	46.04	**51.33**
Masked (COMET)	80.21	89.0	88.97	88.76	71.73	**89.18**	86.93	88.93

Table 4.3.: Results of targeted masking $de \rightarrow en$

	Sentence-Level				Document-Level			
Inference	NN (BL)	NM	MN	MM	NN	NM	MN	MM
Normal (BLEU)	**24.42**	23.56	23.02	23.52	22.58	21.93	21.53	21.91
Normal (COMET)	**81.76**	81.51	81.54	81.15	80.94	80.78	80.73	80.68
Maksed (BLEU)	**22.25**	20.83	20.21	21.03	19.44	19.72	19.48	19.96
Masked (COMET)	79.04	**79.15**	78.92	78.91	77.45	78.61	77.88	78.65

Table 4.4.: Results of targeted masking $en \rightarrow zh$

A similar pattern can also be observed in Chinese→English. Sentence-level models generally achieve better scores than document-level. The SEN-*NM* model performs best in all inference categories.

	Sentence-Level				Document-Level			
Inference	NN (BL)	NM	MN	MM	NN	NM	MN	MM
Normal (BLEU)	16.69	**17.59**	16.58	16.98	15.98	16.84	12.48	16.86
Masked (COMET)	77.71	**78.15**	77.65	77.96	76.28	77.24	75.33	75.83
Normal (BLEU)	11.90	**16.16**	13.69	15.35	12.89	15.00	12.76	15.02
Masked (COMET)	71.93	**76.64**	74.30	76.31	71.11	74.50	71.05	74.14

Table 4.5.: Results of targeted masking $zh \rightarrow en$

Looking generally at the evaluation of the experiments with targeted masking, nothing too surprising can be seen. Models trained on English-Chinese language pair show no tendency of improvements on the document-level side. This could be due to the quite noisy and especially short character of messages uttered in the BMELD conversations. Compared to the synthetic service dialogs in BConTrasT, BMELD provides more informal and realistic dialogs. The English-German pair provides a bit more promising results. Here, document-level models generally achieve better results in masked inference. These find-

ings suggest that contextual information is being used to improve the correct unmasking capability.

4.2.2. Contrastive Evaluation

Following are the results of the contrastive evaluation on the ContraPro dataset. Since ContraPro comprises (*English* → *German*) data only, the corresponding models trained with the targeted masking objective are evaluated here. Similar to the other tables, the character combinations refer to masked or normal data in the respective training stage. Furthermore also scores of models with only first stage training are reported.

It can be seen that DOC-*M* achieves the highest total accuracy, gaining +6 percentage points compared to the second best SL-*M*. Contextual information may play a vital role in the accuracy increase. This supports the thesis that translation of anaphoric pronouns benefits from information in previous utterances.

Also, a strong tendency to translate *it* to *es (n.)* becomes visible. Every model scores higher than 60% accuracy on the neutral pronoun, whereas *er (m.)* and *sie (f.)* scores are significantly lower. DOC-*M* however, achieves the highest accuracy in both non-neutral pronouns, outscoring the next best model by +19 percentage points concerning *er (m.)*. Another observation is the fact that DOC-*M*/DOC-*MM* systems beat unmasked systems DOC-*N*/DOC-*NN*.

Model	Total Accuracy	es (n.)	er (m.)	sie (f.)
DOC_N	0.51	0.68	0.49	0.35
DOC_NN	0.45	0.81	0.23	0.29
DOC_M	0.57	0.65	0.69	0.38
DOC_MM	0.50	0.78	0.45	0.28
SL_N	0.51	0.69	0.50	0.33
SL_NN	0.43	0.81	0.36	0.22
SL_M	0.52	0.77	0.49	0.30
SL_MM	0.45	0.84	0.27	0.23

Table 4.6.: Pronoun Accuracy: Contrastive evaluation results on ContraPro. Note that four exemplary two-stage trained models are shown here to simplify the table. Other two-stage models show similar performance like the ones depicted here (total accuracy ≤ 50% and high accuracy regarding *es*)

Figure 4.1 shows the results of the original study by [26]. Except for the last part of the table, all models are recurrence based. These contain specific architectural changes to comprehend context sentences, like multiple encoders with hierarchical attention (s-hier) and weight tying (s-hier.tied). The s-hier baselines are taken, or slightly adapted from [2], of whom s-hier-to-2 achieves the best accuracy.

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

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

Figure 4.1.: Results from the original ContraPro paper [26].

Since our work did not suggest any structural changes to the network itself, and is based on the Transformer architecture, corresponding results seem more interesting to compare to ours. The tested *concat21* and *concat22* models were taken from [41]. In general, our models show comparable performances regarding total accuracy. They also behave similar concerning the individual pronouns, with a relatively high score for the neutral, and low scores for the gender-specific ones. DOC-*M* outscores the best transformer by +8 points total, achieving worse scores for neutral, outstanding scores for masculine and average scores for feminine pronouns.

These findings provide support for the hypothesis that targeted masking in pretraining is promoting the model's use of contextual information. We used [41] approach to concatenate context, and our masked model outscores all of their context-aware transformers. It remains uncertain specifically how great the targeted approach contributes to the accuracy rise. Future work should therefore examine random masking techniques for comparability. Nonetheless, LM-based masking is preferably selecting words occurring, or being referenced in context sentences. It could be argued, that this is also affecting the results positively.

4.3. Target Context

This section reports results of the experiments with context from target side. The models compared in the following table are elaborated in section **??**. Our proposed Target-Context

model and the basic SL-Baseline are the only "real" NMT models in the table. Others just modify the pronouns of the sentence-level output according to their designation.

Besides the normal BLEU and COMET scores, the share of correctly predicted sentences regarding level of formality is being reported. This pronoun accuracy was calculated manually by comparing each hypothesis sentence to their respective reference in the test split. Table 4.7 shows that *100%-Sie*, *100%-Du* and *Mixed-Sie/Du* models perform similar to the *SL-Baseline* in every reported metric. This is not surprising, as they are just modified versions of the actual sentence-level output. The *100%* models also confirm the expected results regarding pronoun accuracy, since *Formality-BConTrasT* contains approximately half *Sie*-level, half *Du*-level conversations.

The *Target-Context* and *Rule-based* model are indicating a positive correlation between pronoun accuracy and BLEU score. They score about +3 BLEU points better than the baseline, while achieving a notably higher accuracy (around +30 percentage points). The most important finding is that the proposed model acts nearly as good as the assumable perfect *Rule-based* model in pronoun translation, even reaching a slightly better BLEU score.

Model	BLEU	COMET	Pronoun Accuracy
SL-Baseline	48.27	89.2	50.48%
100% Sie	48.49	89.53	50.47%
100% Du	47.9	89.12	49.53%
Mixed Sie/Du	48.28	89.39	49.82%
Target-Context model	51.9	89.19	78.47%
Rule-based model	51.48	89.42	80.1%

Table 4.7.: Results of experiments with target context. *Target-Context* and *Rule-based* model operate with 2 target context sentences.

According to these results, we can infer that contextual information from the target side is beneficial for resolving anaphoric pronouns like "you". Although a simple rule-based model works better in this case, the data supports the idea of utilizing target utterances to gain higher translation quality in conversational settings like chats. For use cases much more complicated than a simple binary decision about formality level, rule-based systems will suffer from the growing dimensionality and won't be able to scale. Therefore NMT models with the ability of handling target context are suggested to succeed in more general experiments.

5. Related Work

Since this work aims at forcing models to exploit contextual information in a chat environment, there are multiple lines of research leading to our experiments. An exemplary list of previous publications is thus given below.

5.1. Conversational Characteristics

Conversational translation aims to capture the nuances and typical discourse phenomena present in chat or dialogue-based interactions. In addition to the rather short and noisy style, existing literature has presented several key characteristics that make conversational data differ from normal text corpuses.

[16] worked out that unlike isolated, unrelated sentences, a discourse normally consists of collocated, structured and coherent groups of sentences. They consider coherence to be a main aspect of discourse-based texts (e.g. dialogs, news-texts, Wikipedia articles).

Another important difference is that conversations not only consist of utterances but also actions. This is stated in [25]. Actions can be questions, made promises, paid compliments and much more. Their study also mentions co-references across several sentences as a substantial part of discourse, and thus conversations.

In their survey paper about DOC-NMT, [24] listed mulitple attributes of discourse data. Among other things, they mention anaphoric pronouns and cohesion. Cohesion refers to the way textual units are linked together grammatically or lexically.

[22] highlights additional chat-related characteristics such as role-preferences (e.g. emotions, style, humor). Different speakers can have different emotions and therefore express themselves differently than others. It was found that explicitly modeling these features through designated latent variables can boost NMT performance over strong baselines.

5.2. Incorporating Context in NMT

Context plays a crucial role in preserving coherence of conversations during translation. In addition to the concatenation approach or masking, which was used in this work, previous papers provide several ways of incorporating context in the NMT workflow modifying model architecture.

Multiple Encoders The first steps were made by adding additional encoders for source-only context. [15] modified the attentional RNN-based NMT architecture with an additional encoder and a corresponding attention element for handling one context sentence. They achieved moderate improvements over small corpuses. To control the information flow from the previous utterance, [19] later used an inter-sentence gate to combine the two context vectors and feed the combination to the decoder. They were able to show that the gate worked effectively in capturing cross sentence-dependencies and lexical cohesion phenomena.

Hierachical Encoders [43] proposed a two-level hierachical RNN to encode the information of three previous source sentences. The resulting vector was then either used to initialize decoder state, as an additional input to decoder state, or after passing through a special context gate. They reported results that surpasses a strong attention-based NMT system by up to +2.1 BLEU points in Chinese-English translation. A combination of hierarchical encoding and the transformer architecture was introduced by [45]. Their proposed encoder first abstracts sentence-level information from context sentences in a transformer-like, self-attentive way, and then hierarchically encodes context-level information. In English-{German, Korean, Turkish} they achieved strong results concerning BLEU scores.

Document-level Training Objectives There has been also more literature trying to implement effective training objectives to exploit contextual information. [31] presented an interesting reinforcement learning approach by utilizing approximated document BLEU as a cost function during training. They demonstrate improvements in English→German for document-level evaluation using TER and BLEU. Recently, [46] proposed the idea of learning "contextualized" embeddings of the source sentence. Therefore, the model was forced to predict the local source context besides the target sentence. These embeddings were then used in a fine-tuning stage. Results better than their transformer baseline were reported concerning Chinese→English and English→German.

6. Conclusion and Future Work

6.1. Conclusion

This study set out to research methods for inducing document-level context awareness in pretrained NMT models for the task of chat translation. Therefore, multiple popular pretrained NMT systems were evaluated on chat data, from whom Facebook's multilingual `M2M100` model (418M parameters) was selected as a starting point for further experiments.

For the purpose of this work, several parallel datasets were found and used, including *News-Commentary*, a non-conversational dataset consisting of news text and commentary. In Addition to two existing chat-specific datasets (BConTrasT, BMELD), a new variation of the BConTrasT data was proposed with modified formality levels and thus, modified pronoun translations.

This research comprised two central experiments, refering to two different training objectives: Targeted Masking as well as target context utilization. For Targeted Masking, the `M2M100` model was trained using 2 consecutive fine-tuning stages, first using non-conversational, and then chat-specific data. Context-agnostic (sentence-level) and context-aware systems were trained, with a context size of 2 previous sentences. Furthermore, masking was applied based on a selection made by a LM, with the aim to mask words that could be resolved by the model using information present in context sentences. All possible combinations of masked and unmasked data were trained, both for English-German and English-Chinese language pairs.

Regarding the second experiment, a document-level model with target context was trained, also with the 2-stage set up described above. For that, the newly created *Formality-BConTrasT* was used during second-stage fine-tuning. The goal was to show that the model can exploit the target context beneficially for predicting the correct German formality level in a conversation. In Addition to a normal sentence-level baseline, several systems with different behaviors regarding pronoun translation were simulated.

Model evaluation was done using automatic metrics BLEU and COMET. In the Targeted Masking setting, no model could outscore the sentence-level baseline concerning BLEU. With masked test data, the document-level model trained with masking in both stages performed best for the English-German language pair. English→German methods were additionally evaluated on a test set with contrastive translations of the anaphoric pronoun it (ContraPro). Here, the document-level model trained only on first-stage masked data achieved notable improvements in pronoun accuracy, even outscoring comparable models from the original ContraPro paper.

The proposed model trained with target context performs slightly better regarding BLEU (+3) and similar regarding COMET than the sentence-level baseline. In a manually

calculated pronoun accuracy score, it nearly reaches the performance of a perfect rule-based model.

In summary, it could be shown that the methods used for training in this work are promoting context-awareness in resulting NMT models. In the case of Targeted Masking, the effect has not been enough to surpass a basic sentence-level model in the chat domain. But results in contrastive evaluation show, that the document-level models are able to resolve typical conversational challenges like anaphoric pronouns. And results with target context show how context can be beneficial in a concrete chat setting.

6.2. Future work

A number of suggestions could be derived from the experiments conducted in this work. A few of them are listed below.

Further context variations Although this work covers a lot of different combinations of context, there are still more than enough for future research. Mixed source and target context could combine the advantages of contextual information from both sides. Proceeding utterances (in settings where the whole chat is available) could be useful to resolve cataphoric pronouns, to just give a few ideas.

New Datasets An NMT model is just as good as the datasets it was trained with are. In a downstream task like chat-translation, the amount and size of available datasets is very limited. Especially datasets with more than two speakers are rare, BMELD being one of few. Therefore, developing new datasets is indispensable when wanting to promote performance of models over time. New datasets should retain a few minimum standards, like short and noisy messages. They should ideally be taken out of real world chat conversations.

More language pairs As with the point above, this is necessary to build a sophisticated translation system for multilingual conversations. In the WMT 2022 shared task on chat translation, English-French and English-Portuguese was added in comparison to WMT 2020. This development is to be supported.

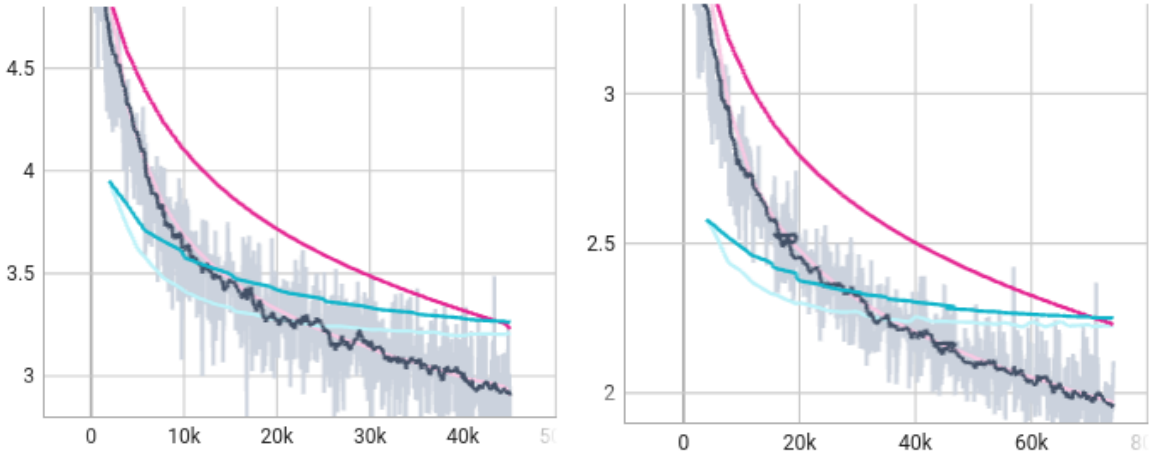
A. Appendix

A.1. Training Params

Here is additional information about the training params we used to train `M2M100` using Fairseq. Parameters slightly differ in patience values regarding training stage. Most of the parameters were initially taken from here and modified by comments of this issue.

A.1.1. First Stage

```
fairseq-train $path_2_data --finetune-from-model $pretrained_model \
--task translation_multi_simple_epoch --encoder-normalize-before \
--save-dir checkpoints_zh_3to1_M_a \
--lang-pairs $lang_pairs --batch-size 4 \
--decoder-normalize-before --sampling-method temperature \
--sampling-temperature 1.5 --encoder-langtok src \
--decoder-langtok --criterion label_smoothed_cross_entropy \
--label-smoothing 0.1 --optimizer adam \
--adam-eps 1e-06 --adam-betas '(0.9, 0.98)' \
--lr-scheduler inverse_sqrt --lr 3e-05 \
--warmup-updates 4000 --dropout 0.3 --weight-decay 0.0 \
--update-freq 2 --validate-interval-updates 5000 \
--no-epoch-checkpoints  --no-last-checkpoints \
--seed 222 --log-interval 10 --patience 5 \
--arch transformer_wmt_en_de_big --encoder-layers 12 --decoder-layers 12 \
--encoder-layerdrop 0.05 --decoder-layerdrop 0.05 \
--share-decoder-input-output-embed \
--share-all-embeddings --ddp-backend no_c10d \
--tensorboard-logdir ./tensorboard_zh_3to1_M_a
```

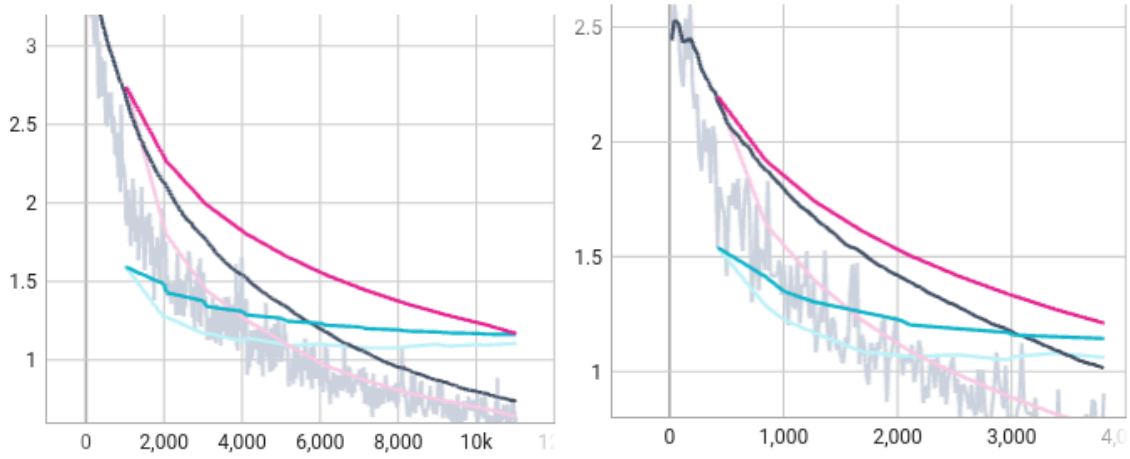


(a) Agent Training (b) Customer Training

Figure A.1.: Exemplary illustrations of the negative log-likelihood loss during the first training stage. Models were trained for 20-25 Epochs, using early stopping after 5 validation turns without loss improvement.

A.1.2. Second Stage

```
fairseq-train $path_2_data --finetune-from-model $pretrained_model \
--task translation_multi_simple_epoch --encoder-normalize-before \
--save-dir ./checkpoints_zh_3to1_MM_a \
--lang-pairs $lang_pairs --batch-size 4 \
--decoder-normalize-before --sampling-method temperature \
--sampling-temperature 1.5 --encoder-langtok src \
--decoder-langtok --criterion label_smoothed_cross_entropy \
--label-smoothing 0.1 --optimizer adam \
--adam-eps 1e-06 --adam-betas '(0.9, 0.98)' \
--lr-scheduler inverse_sqrt --lr 3e-05 \
--warmup-updates 4000 --dropout 0.3 --weight-decay 0.0 \
--update-freq 2 --validate-interval-updates 1000 --no-epoch-checkpoints \
--seed 222 --log-interval 10 --patience 3 \
--arch transformer_wmt_en_de_big --encoder-layers 12 --decoder-layers 12 \
--encoder-layerdrop 0.05 --decoder-layerdrop 0.05 \
--share-decoder-input-output-embed \
--share-all-embeddings --ddp-backend no_c10d \
--tensorboard-logdir ./tensorboard_zh_3to1_MM_a
```



(a) Agent Training

(b) Customer Training

Figure A.1.: Exemplary illustrations of the negative log-likelihood loss during the second training stage. Models were trained for 5-10 Epochs, using early stopping after 3 validation turns without loss improvement.

Bibliography

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[2] Rachel Bawden et al. *Evaluating Discourse Phenomena in Neural Machine Translation*. 2018. arXiv: 1711.00513 [cs.CL].

[3] Bill Byrne et al. *Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset*. 2019. arXiv: 1909.05358 [cs.CL].

[4] Chung-Cheng Chiu et al. "State-of-the-art speech recognition with sequence-to-sequence models". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 4774–4778.

[5] Kyunghyun Cho et al. *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. 2014. arXiv: 1409.1259 [cs.CL].

[6] Ketan Doshi. *Foundations of NLP Explained — Bleu Score and WER Metrics*. URL: https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b (visited on 05/09/2021).

[7] Pranay Dugar. *Attention - Seq2Seq Models*. URL: https://towardsdatascience.com/day-1-2-attention-seq2seq-models-65df3f49e263 (visited on 07/13/2019).

[8] Angela Fan et al. "Beyond English-Centric Multilingual Machine Translation". In: *J. Mach. Learn. Res.* 22.1 (Jan. 2021). ISSN: 1532-4435.

[9] M. Amin Farajian et al. "Findings of the WMT 2020 Shared Task on Chat Translation". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 65–75. URL: https://aclanthology.org/2020.wmt-1.3.

[10] Ana C Farinha et al. "Findings of the WMT 2022 Shared Task on Chat Translation". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 724–743. URL: https://aclanthology.org/2022.wmt-1.70.

[11] Philip Gage. "A new algorithm for data compression". In: *C Users Journal* 12.2 (1994), pp. 23–38.

[12] Jonas Gehring et al. "Convolutional Sequence to Sequence Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1243–1252. URL: https://proceedings.mlr.press/v70/gehring17a.html.

[13] Frauke Günther and Stefan Fritsch. "Neuralnet: training of neural networks." In: *R J.* 2.1 (2010), p. 30.

[14] Christian Hardmeier et al. "Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation". In: *Second Workshop on Discourse in Machine Translation (DiscoMT), 17 September 2015, Lisbon, Portugal*. Association for Computational Linguistics. 2015, pp. 1–16.

[15] Sebastien Jean et al. *Does Neural Machine Translation Benefit from Larger Context?* 2017. arXiv: 1704.05135 [stat.ML].

[16] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[17] Chetna Khanna. *Byte-Pair Encoding: Subword-based tokenization algorithm*. URL: https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0 (visited on 08/13/2021).

[18] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proc. ACL*. 2017. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012.

[19] Shaohui Kuang and Deyi Xiong. "Fusing Recency into Neural Machine Translation with an Inter-Sentence Gate Model". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 607–617. URL: https://aclanthology.org/C18-1051.

[20] Taku Kudo. *Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates*. 2018. arXiv: 1804.10959 [cs.CL].

[21] Chia-Hsuan Lee et al. *DOCmT5: Document-Level Pretraining of Multilingual Language Models*. 2022. arXiv: 2112.08709 [cs.CL].

[22] Yunlong Liang et al. "Modeling Bilingual Conversational Characteristics for Neural Chat Translation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 5711–5724. DOI: 10.18653/v1/2021.acl-long.444. URL: https://aclanthology.org/2021.acl-long.444.

[23] Yinhan Liu et al. *Multilingual Denoising Pre-training for Neural Machine Translation*. 2020. arXiv: 2001.08210 [cs.CL].

[24] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. "A Survey on Document-Level Neural Machine Translation: Methods and Evaluation". In: *ACM Comput. Surv.* 54.2 (Mar. 2021). ISSN: 0360-0300. DOI: 10.1145/3441691. URL: https://doi.org/10.1145/3441691.

[25] Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*. Vol. 6. 94. Springer, 2016.

[26] Mathias Müller et al. "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, 2018.

[27] Nathan Ng et al. *Facebook FAIR's WMT19 News Translation Task Submission*. 2019. arXiv: 1907.06616 [cs.CL].

[28] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[29] Matt Post and Marcin Junczys-Dowmunt. “Escaping the sentence-level paradigm in machine translation”. In: *arXiv preprint arXiv:2304.12959* (2023).

[30] Ricardo Rei et al. *COMET: A Neural Framework for MT Evaluation*. 2020. arXiv: 2009.09025 [cs.CL].

[31] 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*. Online: Association for Computational Linguistics, July 2020, pp. 7764–7770. DOI: 10.18653/v1/2020.acl-main.693. URL: https://aclanthology.org/2020.acl-main.693.

[32] Holger Schwenk. “Continuous space translation models for phrase-based statistical machine translation”. In: *Proceedings of COLING 2012: Posters*. 2012, pp. 1071–1080.

[33] Holger Schwenk, Daniel Déchelotte, and Jean-Luc Gauvain. “Continuous space language models for statistical machine translation”. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. 2006, pp. 723–730.

[34] 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)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

[35] Sagar Sharma. *What the Hell is Perceptron?* URL: https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53 (visited on 09/11/2017).

[36] Yusuxke Shibata et al. “Byte Pair encoding: A text compression scheme that accelerates pattern matching”. In: (1999).

[37] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems* 27 (2014).

[38] Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. “A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization”. In: *IEEE Access* 9 (2021), pp. 13248–13265. DOI: 10.1109/ACCESS.2021.3052783.

[39] Chuanqi Tan et al. “A Survey on Deep Transfer Learning”. In: *Artificial Neural Networks and Machine Learning – ICANN 2018*. Ed. by Věra Kůrková et al. Cham: Springer International Publishing, 2018, pp. 270–279. ISBN: 978-3-030-01424-7.

[40] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012. ISBN: 978-2-9517408-7-7.

[41] Jörg Tiedemann and Yves Scherrer. "Neural Machine Translation with Extended Context". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 82–92. DOI: 10.18653/v1/W17-4811. URL: https://aclanthology.org/W17-4811.

[42] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[43] Longyue Wang et al. "Exploiting Cross-Sentence Context for Neural Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2826–2831. DOI: 10.18653/v1/D17-1301. URL: https://aclanthology.org/D17-1301.

[44] Jonathan J Webster and Chunyu Kit. "Tokenization as the initial phase in NLP". In: *COLING 1992 volume 4: The 14th international conference on computational linguistics*. 1992.

[45] Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. "Improving Context-Aware Neural Machine Translation Using Self-Attentive Sentence Embedding". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 9498–9506. DOI: 10.1609/aaai.v34i05.6494. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6494.

[46] Pei Zhang et al. *Learning Contextualized Sentence Representations for Document-Level Neural Machine Translation*. 2020. arXiv: 2003.13205 [cs.CL].