
Proceedings Paper:
Vincent, S., Flynn, R. and Scarton, C. orcid.org/0000-0002-0103-4072 (2023) MTCue: learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In: Findings of the Association for Computational Linguistics: ACL 2023. Findings of the Association for Computational Linguistics: ACL 2023, 09-14 Jul 2023, Toronto, Canada. Association for Computational Linguistics , pp. 8210-8226. ISBN 9781959429623

https://doi.org/10.18653/v1/2023.findings-acl.521

MTCue: Learning Zero-Shot Control of Extra-Textual Attributes by Leveraging Unstructured Context in Neural Machine Translation

Sebastian Vincent and **Robert Flynn** and **Carolina Scarton**
Department of Computer Science, University of Sheffield, UK
{stvincent1,rjflynn2,c.scarton}@sheffield.ac.uk

Abstract

Efficient utilisation of both intra- and extra-textual context remains one of the critical gaps between machine and human translation. Existing research has primarily focused on providing individual, well-defined types of context in translation, such as the surrounding text or discrete external variables like the speaker's gender. This work introduces MTCue, a novel neural machine translation (NMT) framework that interprets all context (including discrete variables) as text. MTCue learns an abstract representation of context, enabling transferability across different data settings and leveraging similar attributes in low-resource scenarios. With a focus on a dialogue domain with access to document and metadata context, we extensively evaluate MTCue in four language pairs in both translation directions. Our framework demonstrates significant improvements in translation quality over a parameter-matched non-contextual baseline, as measured by BLEU (+0.88) and COMET (+1.58). Moreover, MTCue significantly outperforms a "tagging" baseline at translating English text. Analysis reveals that the context encoder of MTCue learns a representation space that organises context based on specific attributes, such as formality, enabling effective zero-shot control. Pre-training on context embeddings also improves MTCue's few-shot performance compared to the "tagging" baseline. Finally, an ablation study conducted on model components and contextual variables further supports the robustness of MTCue for context-based NMT.

github.com/st-vincent1/MTCue

1 Introduction

Research in neural machine translation (NMT) has advanced considerably in recent years, much owing to the release of the Transformer architecture (Vaswani et al., 2017), subword segmentation (Sennrich et al., 2016c) and back-translation (Sennrich et al., 2016b). This resulted in claims of human parity in machine translation (Hassan et al., 2018), which in turn prompted researchers to look beyond the sentence level: at how a translation still needs to be compatible with the context it arises in.

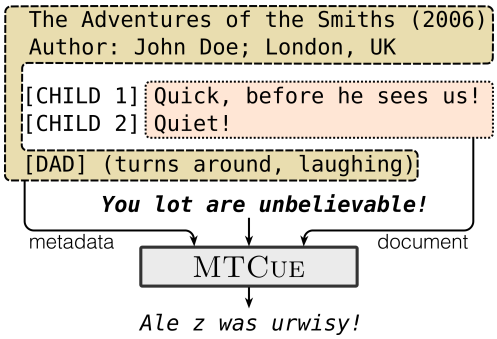


Figure 1: A high-level overview of MTCue (EN→PL).

The task of contextual adaptation to more nuanced extra-textual variables like the description of the discourse situation has been largely overlooked, in spite of earlier work suggesting that conversational machine translation may benefit from such fine-grained adaptations (van der Wees et al., 2016). Most existing work on contextual NMT has focused on document-level context instead, aiming to improve the coherence and cohesion of the translated document (e.g. Tiedemann and Scherrer, 2017). Some research has successfully adapted NMT to extra-textual context variables using supervised learning frameworks on labelled datasets, targeting aspects such as gender (Vanmassenhove et al., 2018; Moryossef et al., 2019; Vincent et al., 2022b), formality (Sennrich et al., 2016a; Nadejde et al., 2022), translators' or speakers' style (Michel and Neubig, 2018a; Wang et al., 2021b) and translation length (Lakew et al., 2019), sometimes controlling multiple attributes simultaneously (Schioppa et al., 2021; Vincent et al., 2022b). However, to our knowledge, no prior work has attempted to model the impact of continuous extra-textual contexts in translation or combined the intra- and extra-textual contexts in a robust framework. This is

Findings of the Association for Computational Linguistics: ACL 2023, pages 8210–8226
July 9-14, 2023

problematic since translating sentences without or with incomplete context is akin to a human translator working with incomplete information. Similarly, only a handful of earlier research has contemplated the idea of controlling these extra-textual attributes in a zero-shot or few-shot fashion (Moryossef et al., 2019; Anastasopoulos et al., 2022); such approaches are essential given the difficulty of obtaining the labels required for training fully supervised models.

In some domains, extra-textual context is paramount and NMT systems oblivious to this information are expected to under-perform. For instance, for the dubbing and subtitling domain, where translated shows can span different decades, genres, countries of origin, etc., a one-size-fits-all model is limited by treating all input sentences alike. In this domain, there is an abundance of various metadata (not just document-level data) that could be used to overcome this limitation. However, such adaptation is not trivial: (i) the metadata often comes in quantities too small for training and with missing labels; (ii) it is expressed in various formats and types, being difficult to use in a standard pipeline; (iii) it is difficult to quantify its exact (positive) effect.

In this paper, we address (i) and (ii) by proposing MTCUE (**M**achine **T**ranslation with **C**ontextual **u**niversal **e**mbeddings), a novel NMT framework that bridges the gap between training on discrete control variables and intra-textual context as well as allows the user to utilise metadata of various lengths in training, easing the need for laborious data editing and manual annotation (Figure 1). During inference, when context is provided verbatim, MTCUE falls back to a code-controlled translation model; by vectorising the inputs, it exhibits competitive performance for noisy phrases and learns transferrability across contextual tasks. While (iii) is not directly addressed, our evaluation encompasses two translation quality metrics and two external test sets of attribute control, showing the impact on both translation quality and capturing relevant contextual attributes.

MTCUE can generalise to unseen context variables, achieving 100% accuracy at a zero-shot formality controlling task; it learns to map embeddings of input contexts to discrete phenomena (e.g. formality), increasing explainability; and it exhibits more robust few-shot performance at multi-attribute control tasks than a "tagging" baseline.

The main contributions of this work are:

1. MTCUE (§2): a novel framework for **combining (un)structured intra- and extra-textual context in NMT** that significantly improves translation quality for four language pairs in both directions: English (EN) to/from German (DE), French (FR), Polish (PL) and Russian (RU).
2. A comprehensive evaluation, showing that MTCUE can be primed to exhibit **excellent zero-shot and few-shot performance** at downstream contextual translation tasks (§4 and §5).
3. Pre-trained models, code, and an organised version of the OpenSubtitles18 (Lison et al., 2018) dataset **with the annotation of six metadata** are made available.

This paper also presents the experimental settings (§3), related work (§6) and conclusions (§7).

2 Proposed Architecture: MTCUE

MTCUE is an encoder-decoder Transformer architecture with two encoders: one dedicated for contextual signals and one for inputting the source text. The signals from both encoders are combined using parallel cross-attention in the decoder. Below we describe how context inputs are treated in detail, and later in §2.2 and §2.3 we describe the context encoder and context incorporation, respectively.

2.1 Vectorising Contexts

Context comes in various formats: for example, the speaker's gender or the genre of a film are often supplied in corpora as belonging to sets of predetermined discrete classes, whereas plot descriptions are usually provided as plain text (and could not be treated as discrete without significant loss of information). To leverage discrete variables as well as short and long textual contexts in a unified framework, we define a **vectorisation function** that maps each context to a single meaningful vector, yielding a matrix $\mathbf{E}_{c\times r}$, where c is the number of contexts and r is the embedding dimension. The function is deterministic (the same input is always embedded in the same way) and semantically coherent (semantically similar inputs receive similar embeddings). We use a sentence embedding model (Reimers and Gurevych, 2019) for vectorisation, which produces embeddings both deterministic and semantically coherent. Motivated by Khandelwal et al. (2018) and O'Connor and Andreas (2021) who report that generation models mostly use general topical information from past context, ignor-

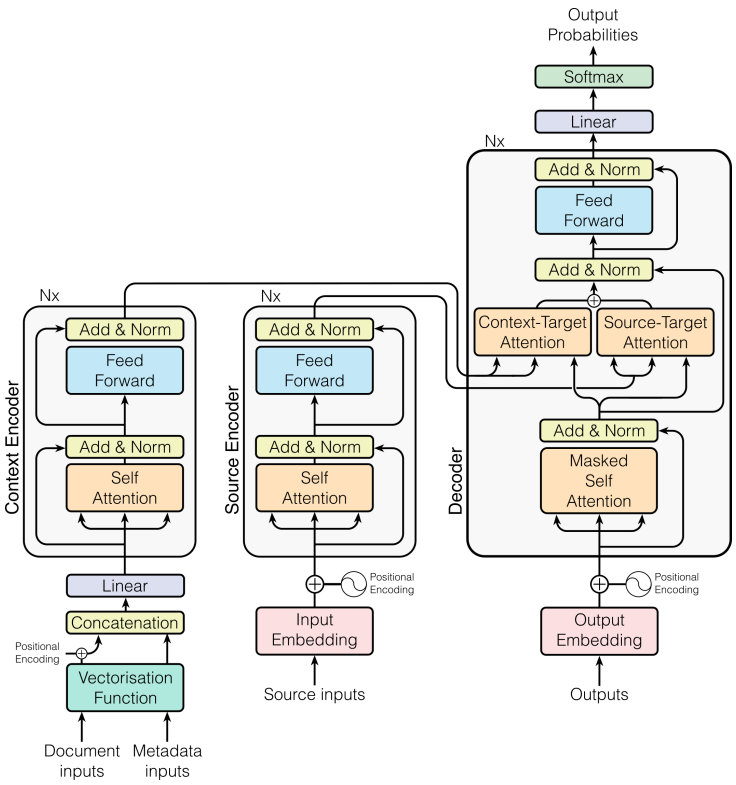


Figure 2: The MTCUE architecture. Stylised after the Transformer architecture figure in (Vaswani et al., 2017).

ing manipulations such as shuffling or removing non-noun words, we hypothesise that sentence embeddings can effectively compress the relevant context information into a set of vectors, which, when processed together within a framework, will formulate an abstract representation of the dialogue context. We select the MINILMV2 sentence embedding model (Wang et al., 2021a), which we access via the `sentence-transformers` library;[1] a similar choice was made concurrently in Vincent et al. (2023). In the experiments, we also refer to DISTILBERT (Sanh et al., 2019) which is used by one of our baselines, and a discrete embedding function which maps unique contexts to the same embeddings but has no built-in similarity feature.

For any sample, given a set of its k textual contexts $C = [c_1, ...c_k]$, we vectorise each one separately using the method described above. The resulting array of vectors is the input we supply to the context encoder in MTCUE.

2.2 Context Encoder

Processing vectorised contexts The context encoder of MTCUE is a standard self-attention encoder with a custom input initialisation. Its inputs are sentence embeddings of context (§2.1) projected to the model's dimensions with a linear layer ($384 \rightarrow d_{model}$). In preliminary experiments, we observe that the first layer of the context encoder receives abnormally large input values, which sometimes leads to the explosion of the query ($\mathbf{Q}$) and key ($\mathbf{K}$) dot product $\mathbf{QK}^{\mathrm{T}}$. We prevent this by replacing the scaled dot product attention with query-key normalisation (Henry et al., 2020): applying L2 normalisation of $\mathbf{Q}$ and $\mathbf{K}$ before the dot product, and replacing the scaling parameter $\sqrt{d}$ with a learned one, initialised to a value based on training data lengths.[2]

Positional embeddings We use positional context embeddings to (a) indicate the distance of a past utterance to the source sentence and (b) to distinguish metadata inputs from document information. In particular, when translating the source sentence s_i at position i in the document, a sentence distance positional embedding (POS) is added to the embedding representations of each past sentence s_{i-j}, with $j \in [0, t]$ where t is the maximum allowed context distance: $e'(s_{i-j}, j) = e(s_{i-j}) + POS(j)$. Metadata contexts $(m_0, \ldots, m_n)$ do not receive positional em-

[1] https://sbert.net/, accessed 1/5/23.

[2] An alternative solution applies layer normalisation to the input of the first layer, but we found that this degraded performance w.r.t. QK-NORM.

beddings since their order is irrelevant. The final vectorised input of the context encoder is: $e'(s_i, 0), e'(s_{i-1}, 1), \ldots, e'(s_{i-t}, t), e(m_0), \ldots, e(m_n)$.

2.3 Context incorporation

The outputs of the context and source encoders (respectively $\mathcal{C}$ and $\mathcal{S}$) are combined in the decoder using **parallel attention** (Libovický et al., 2018). Let the output of the decoder self-attention be $\mathcal{T}$. Let $\mathcal{T}_{out} = \text{FFN}(\mathcal{T}') + \mathcal{T}'$, where $\mathcal{T}'$ is the multi-head attention output; i.e. $\mathcal{T}_{out}$ is $\mathcal{T}'$ with the feed-forward layer and the residual connection applied. In a non-contextual Transformer, source and target representations are combined with cross-attention:

$$\mathcal{T}' = \text{mAttn}(kv = \mathcal{S}, q = \mathcal{T})$$

In contrast, parallel attention computes individual cross-attention of $\mathcal{T}$ with $\mathcal{S}$ and $\mathcal{C}$ and then adds them together:

$$\mathcal{S}' = \text{mAttn}(kv = \mathcal{S}, q = \mathcal{T})$$
$$\mathcal{C}' = \text{mAttn}(kv = \mathcal{C}, q = \mathcal{T})$$
$$\mathcal{T}' = \mathcal{C}' + \mathcal{S}'$$

Parallel attention is only one of many combination strategies which can be used, and in preliminary experiments we found the choice of the strategy to have a minor impact on performance.

3 Experimental Setup

3.1 Data: the OpenSubtitles18 Corpus

Data type	EN↔DE	EN↔FR	EN↔PL	EN↔RU
Source & target	5.3M	14.7M	12.9M	12.4M
metadata				
Genre	45.3%	57.8%	60.5%	73.4%
PG rating	35.9%	46.9%	48.8%	62.3%
Writer(s)	45.3%	57.1%	58.9%	71.7%
Year	45.3%	57.8%	60.5%	73.7%
Country	37.7%	42.9%	45.7%	42.7%
Plot description	43.4%	57.1%	59.7%	72.6%
previous dialogue				
$n-1$	60.1%	68.0%	63.7%	73.6%
$n-2$	42.0%	51.2%	46.4%	57.9%
$n-3$	31.2%	40.1%	35.5%	46.9%
$n-4$	23.9%	32.2%	28.0%	38.6%
$n-5$	18.7%	26.2%	22.4%	32.2%

Table 1: Data quantities for the extracted OpenSubtitles18 corpus. An average of 81% samples has at least one context input.

The publicly available OpenSubtitles18[3] corpus (Lison et al., 2018), hereinafter OPENSUBTITLES, is a subtitle dataset in `.xml` format with IMDb ID attribution and timestamps. It is a mix of original and user-submitted subtitles for movies and TV content. Focusing on four language pairs (EN↔{DE,FR,PL,RU}), we extract parallel sentence-level data with source and target document-level features (up to 5 previous sentences) using the timestamps (see Appendix A). We also extract a range of metadata by matching the IMDb ID against the Open Movie Database (OMDb) API.[4] Table 1 shows training data quantities and portions of annotated samples per context while Table 2 shows an example of the extracted data. We select six metadata types that we hypothesise to convey useful extra-textual information: *plot description* (which may contain useful topical information), *genre* (which can have an impact on the language used), *year of release* (to account for the temporal dimension of language), *country of release* (to account for regional differences in expression of English), *writers* (to consider writers' style), *PG rating* (which may be associated with e.g. the use of adult language). For validation and testing, we randomly sample 10K sentence pairs each from the corpus, based on held-out IMDb IDs.

Key	Value
Source (EN)	This is the Angel of Death, big daddy reaper.
Target (PL)	To anioł śmierci. Kosiarz przez wielkie "k".
PG rating	PG rating: TV-14
Released	Released in 2009
Writers	Writers: Eric Kripke, Ben Edlund, Julie Siege
Plot	Dean and Sam get to know the whereabouts of Lucifer and want to hunt him down. But Lucifer is well prepared and is working his own plans.
Genre	Drama, Fantasy, Horror
Country	United States, Canada

Table 2: Example of a source-target pair and metadata in OPENSUBTITLES.

Preprocessing The corpus is first detokenised and has punctuation normalised (using Moses scripts (Koehn et al., 2007)). Then a custom cleaning script is applied, which removes trailing dashes, unmatched brackets and quotation marks, and fixes common OCR spelling errors. Finally, we perform subword tokenisation via the BPE algorithm with Sentencepiece (Kudo and Richardson, 2018).

Film metadata (which comes from OMDb) is left intact except when the fields contain non-values such as "N/A", "Not rated", or if a particular field is not sufficiently descriptive (e.g. a PG rating field represented as a single letter "R"), in which case

[3] Created from data from `https://opensubtitles.org`.

[4] `https://omdbapi.com/`, accessed 1/5/23.

we enrich it with a disambiguating prefix (e.g. "R" → "PG rating: R"). Regardless of the trained language pair, metadata context is provided in English (which here is either the source or target language). Document-level context is limited to source-side context. Since for *→EN language pairs the context input comes in two languages (e.g. English metadata and French dialogue), we use multilingual models to embed the context in these pairs.

3.2 Evaluation

We evaluate the presented approach with the general in-domain test set as well as two external contextual tasks described in this section.

Translation quality The approaches are evaluated against an in-domain held-out test set of 10K sentence pairs taken from OPENSUBTITLES. As metrics, we use BLEU[5] (Papineni et al., 2002) and COMET[6] (Rei et al., 2020).

Control of multiple attributes about dialogue participants (EAMT22) The EAMT22 task, introduced by Vincent et al. (2022b), evaluates a model's capability to control information about dialogue participants in English-to-Polish translation. The task requires generating hypotheses that align with four attributes: gender of the speaker and interlocutor(s) (masculine/feminine/mixed), number of interlocutors (one/many), and formality (formal/informal). These attributes can occur in a total of 38 unique combinations. We investigate whether MTCUE can learn this task through zero-shot learning (pre-training on other contexts) or through few-shot learning (when additionally fine-tuned on a constrained number of samples).

To prepare the dataset, we use scripts provided by Vincent et al. (2022b) to annotate OPENSUBTITLES with the relevant attributes, resulting in a corpus of 5.1M annotated samples. To leverage the context representation in MTCUE, we transcribe the discrete attributes to natural language by creating three sentences that represent the context. For example, if the annotation indicates that the speaker is male, the interlocutor is a mixed-gender group, and the register is formal, we create the following context: (1) "I am a man", (2) "I'm talking to a group of people" and (3) "Formal".

We train seven separate instances of MTCUE using different artificial data settings. Each setting contains the same number of samples (5.1M) but a varying number of **annotated** samples. To address class imbalances in the dataset (e.g. *masculine speaker* occurring more often than *feminine speaker*) and ensure equal representation of the 38 attribute combinations, we collect multiples of these combinations. We select sample numbers to achieve roughly equal logarithmic distances: 1, 5, 30, 300, 3K and 30K supervised samples per each of 38 combinations, yielding exactly 38, 180, 1,127, 10,261, 81,953 and 510,683 samples respectively. Including the zero-shot and full supervision (5.1M cases), this results in a total of eight settings. Each model is trained with the same hyperparameters as MTCUE, and on the same set of 5.1M samples, with only the relevant number of samples annotated (non-annotated samples are given as source-target pairs without contexts). We compare our results against our re-implementation of the TAGGING approach which achieved the best performance in the original paper (i.e. Vincent et al., 2022b). We train the TAGGING model in replicas of the eight settings above.

Zero-shot control of formality (IWSLT22) We experiment with the generalisation of MTCUE to an unseen type of context: formality. In the IWSLT22 formality control task (Anastasopoulos et al., 2022), the model's challenge is to produce hypotheses agreeing with the desired formality (formal/informal). For the English-to-German language pair, the task provides a set of paired examples (each source sentence is paired with a formal reference and an informal one), to a total of 400 validation and 600 test examples; for the English-to-Russian pair, only the 600 test examples are provided. We test the capacity of MTCUE to control formality zero-shot, given a textual cue as context input.[7]

3.3 Baselines

In our experiments, we compare MTCUE with three types of baselines:

1. **BASE and BASE-PM**. These are pre-trained translation models that match MTCUE either in the shape of the encoder-decoder architecture (BASE) or in terms of the total number of parameters (BASE-PM). For BASE-PM, the extra parameters are obtained from enhancing the source encoder, increasing the number

[5]Computed with SacreBLEU (Post, 2018).
[6]Computed using the `wmt20-comet-da` model.
[7]We describe the process of choosing the context input for evaluation in Appendix D.

Model	Params	d_{model}	Layers			h	FFN dim.			GPU Hour/Epoch	Epochs to best
			Cxt	Src	Dec		Cxt	Src	Dec		
BASE	66M	512	–	6	6	8	–	2048	2048	–	–
BASE-PM	107M	512	–	10	6	8	–	4096	2048	–	–
TAGGING	107M	512	–	10	6	8	–	4096	2048	0.74 ± 0.35	6.13 ± 4.09
NOVOTNEY-CUE	99M	512	6	6	6	8	2048	2048	2048	1.29 ± 0.56	9.13 ± 3.60
MTCUE	105M	512	6	6	6	8	2048	2048	2048	0.81 ± 0.39	9.38 ± 4.57

Table 3: Model details for MTCUE and baselines. Timings and epochs are averaged across all language directions.

of layers ($6 \rightarrow 10$) and doubling the feed-forward dimension ($2048 \rightarrow 4096$).

2. **TAGGING**. Following previous work (e.g. Schioppa et al., 2021; Vincent et al., 2022b), we implement a model that assigns a discrete embedding to each unique context value. Architecturally, the model matches BASE-PM. The tags are prepended to feature vectors from the source context and then together fed to the decoder.

3. **NOVOTNEY-CUE**. This baseline is a re-implementation of the CUE vectors architecture (Novotney et al., 2022) for NMT. It utilises DISTILBERT for vectorisation and averages the context feature vectors to obtain the decoder input. In contrast, MTCUE employs a parallel attention strategy.

In experiments on formality control, we also report results from the two submissions to the IWSLT22 task, both implementing a supervised and a zero-shot approach:

1. Vincent et al. (2022a). This (winning) submission combines the TAGGING approach with formality-aware re-ranking and data augmentation. The authors augment the original formality-labelled training samples by matching sentence pairs from larger corpora against samples of specific formality (akin to the Moore-Lewis algorithm described in Moore and Lewis, 2010). Their zero-shot approach relies on heuristically finding a suitable sample of formality-annotated data similar to the provided set and performing the same algorithm above.

2. Rippeth et al. (2022) who fine-tune large pre-trained multilingual MT models with additive control (Schioppa et al., 2021) on data with synthetic formality labels obtained via rule-based parsers and classifiers.

3.4 Implementation and hyperparameters

We implement MTCUE and all its components in FAIRSEQ, and use HuggingFace (Wolf et al., 2020) for vectorising contexts. We use hyperparameters recommended by FAIRSEQ, plus optimise the learning rate and the batch size in a grid search. We found that a learning rate of 0.0003 and a batch size of simulated 200K tokens worked best globally. Table 3 presents the architecture details and runtimes for the models. All training is done on a single A100 80GB GPU, one run per model. We use early stopping based on validation loss with a patience of 5.

4 Results

Translation quality Results in Table 4 show that MTCUE beats all non-contextual baselines in translation quality, achieving an average improvement of $+1.51$ BLEU/$+3.04$ COMET over BASE and $+0.88$/$+1.58$ over BASE-PM. It is also significantly better than NOVOTNEY-CUE ($+0.46$/$+0.66$). MTCUE achieves comparable results to the parameter-matched TAGGING model, consistently outperforming it on all language directions from English, and being outperformed by it on directions into English. Since the primary difference between the two models is that MTCUE sacrifices more parameters to process context, and TAGGING uses these parameters for additional processing of source text, we hypothesise that the difference in scores is due to the extent to which context is a valuable signal for the given language pair: it is less important in translation into English. This is supported by findings from literature: English is a language that does not grammatically mark phenomena such as gender (Stahlberg et al., 2007).

The largest quality improvements with MTCUE are obtained on EN-DE ($+1.66$/$+4.14$ vs BASE-PM and $+1.14$/$+1.70$ vs TAGGING) and EN-FR ($+2.23$/$+3.32$ vs BASE-PM and $+0.80$/$+0.62$ vs TAGGING) language pairs. Contrastively, the smallest improvements against BASE-PM are obtained on the RU-EN pair. MTCUE is outperformed by

Model	EN→DE		EN→FR		EN→PL		EN→RU		DE→EN		FR→EN		PL→EN		RU→EN		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Baselines																		
*BASE	33.60	45.90	34.54	46.92	28.08	58.52	31.37	62.94	39.53	59.56	35.46	55.10	34.42	50.38	39.37	55.99	34.65	54.41
*BASE-PM	34.36	46.77	35.31	48.87	28.66	60.97	32.40	64.55	40.32	60.88	36.16	56.28	35.03	51.77	40.04	56.86	35.28	55.87
TAGGING	34.88	49.21	36.74	51.57	29.08	**64.29**	32.32	65.12	**41.52**	**62.63**	**37.10**	**57.41**	**36.19**	**53.46**	**40.33**	**57.14**	36.02	**57.60**
NOVOTNEY-CUE	35.30	49.83	36.75	50.52	29.09	62.69	32.36	64.90	40.86	61.91	36.51	56.21	35.28	52.17	39.44	56.08	35.70	56.79
Proposed																		
MTCUE	**36.02**	**50.91**	**37.54**	**52.19**	**29.36**	63.46	**33.21**	**65.21**	40.95	61.58	36.57	56.87	35.68	52.48	39.97	56.92	**36.16**	57.45

Table 4: Translation quality results on the OPENSUBTITLES test set. *Model trained without access to any context. We highlight the best result in each column and underline all statistically indistinguishable results, $p \leq 0.05$ (except the Average column).

TAGGING the most on PL-EN ($-0.51/-0.98$). As far as training efficiency, MTCUE trains significantly faster than NOVOTNEY-CUE, converging in a similar number of epochs but using significantly less GPU time, on par with TAGGING (Table 3). Finally, all contextual models considered in this evaluation significantly outperform the parameter-matched translation model (BASE-PM), clearly signalling that metadata and document context are an important input in machine translation within this domain, regardless of the chosen approach.

Control of multiple attributes about dialogue participants (EAMT22) MTCUE achieves 80.25 zero-shot accuracy at correctly translating the speaker and interlocutor attributes, an improvement of 12.08 over the non-contextual baseline, also expressed in increased translation quality (25.22 vs 23.36 BLEU). Furthermore, it bests TAGGING at few-shot performance by 5 to 8 accuracy points, reaching above 90% accuracy with only 190 of the 5.1M annotated samples (Figure 4). Both TAGGING and MTCUE perform similarly with more supervised data. The TAGGING model achieves +2 to +3 accuracy points in the 1K to 100K range, while BLEU remains comparable. We hypothesise that this happens because MTCUE relies strongly on its pre-training prior when context is scarce: this proves useful with little data, but becomes less relevant as more explicitly labelled samples are added. Finally, with full supervision, both models achieve above 99% accuracy.

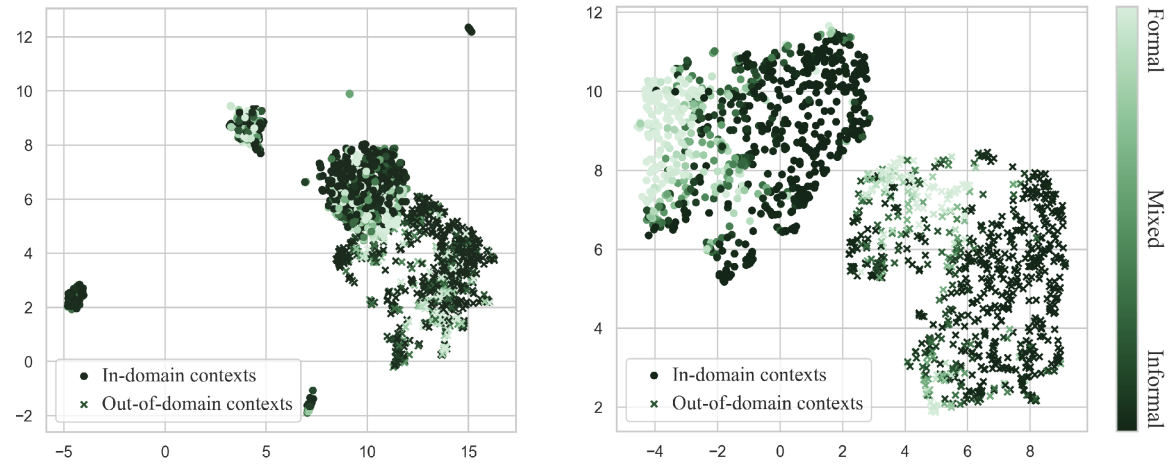
Zero-shot control of formality (IWSLT22) MTCUE appears to successfully control the formality of translations in a zero-shot fashion, achieving nearly 100% accuracy on the IWSLT22 test sets across two language pairs, beating all zero-shot models on the EN-RU pair and performing on par with the best supervised model for EN-DE. Notably, both baselines presented in Table 5 were built to target formality specifically, unlike MTCUE which is a general-purpose model.

	Model	Supervision	Formal	Informal	Average
EN-DE	Non-context baseline	–	74.5	25.5	50.0
	Rippeth et al. (2022)	Supervised	99.4	96.5	98.0
	Vincent et al. (2022a)	Supervised	100.0	100.0	**100.0**
	MTCUE	Zero-shot	100.0	100.0	**100.0**
EN-RU	Non-context baseline	–	96.4	3.6	50.0
	Rippeth et al. (2022)	Zero-shot	100.0	1.1	50.5
	Vincent et al. (2022a)	Zero-shot	99.5	85.8	92.7
	MTCUE	Zero-shot	100.0	99.4	**99.7**

Table 5: Evaluation on the IWSLT22 formality control evaluation campaign. Baseline systems were trained on different corpora.

Following MTCUE's success at controlling formality with sample contexts, we investigate the relationship between context embeddings and their corresponding formality control scores. We consider all 394 unique contexts from the OPENSUBTITLES validation data, and another 394 document contexts (individual past sentences) at random (in-domain). We also use an in-house dataset from a similar domain (dubbing of reality cooking shows with custom annotations of scene contents) and select another 394 metadata and 394 document contexts from there (out-of-domain). We run inference on the IWSLT22 test set with each context individually (1,576 runs), and use UMAP (McInnes et al., 2018) to visualise (i) the input embedding from MINILM-V2, (ii) the output vector of the context encoder and (iii) the corresponding formality score (Figure 3).

We invite the reader to pay attention to the separation of dark and light points in Figure 3b that is not present in Figure 3a. There is a spatial property that arises in the context encoder and is shown by Figure 3b, namely a relationship between the feature vectors from context encoder and formality scores across both domains: contexts yielding translations of the same register tend to be clustered together. This is true for both in-domain data (cir-



(a) MINILM-V2 embeddings. (b) Output of MTCUE's context encoder.

Figure 3: UMAP visualisation of how various contexts impact the formality of produced translations when used as input in MTCUE.

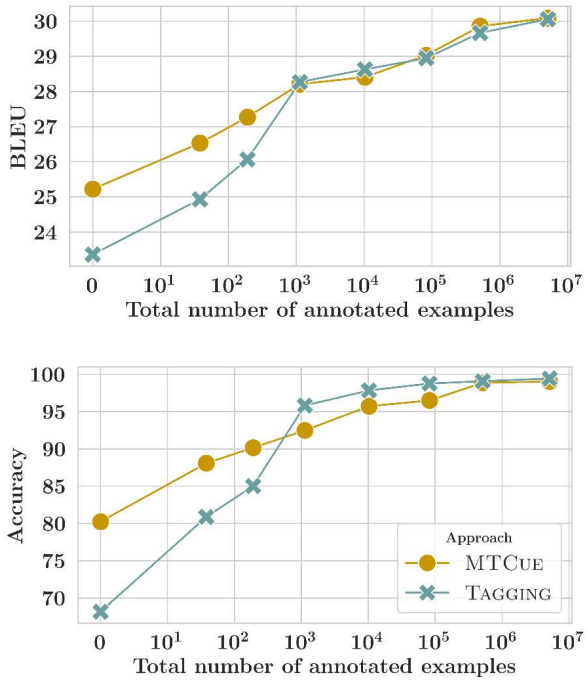


Figure 4: Evaluation results from the EAMT22 multi-attribute control task.

cles) and out-of-domain data (crosses), suggesting that after training this effect generalises to unseen contexts.

For further investigation, we sample a few contexts at random which yield 100% zero-shot accuracy (from the "ends" of the color scale) and find that these contexts tend to have semantic relationships with the type of formality they induce in translations. For example, contexts like "What's wrong with you?", "Wh-what's he doing now?" yield all-informal translations while "Then why are you still in my office?" or "I can see you're very interested." result in all-formal ones. This confirms our hypothesis: MTCUE's context encoder aligns the semantic representation of the input context to the most likely formality it would produce, akin to a human translator deducing such information from available data. Outside of an evaluation scenario like the present one, MTCUE may therefore be able to predict from the given context what formality style should be used: an effect only facilitated by the context encoder.

To exemplify how the zero-shot performance of MTCUE manifests in practice, we present some examples of outputs for the two tasks in Appendix E.

5 Ablation Study

Ablation	COMET EN→DE	COMET EN→FR	COMET EN→PL	ZERO-SHOT ACCURACY IWSLT22 (DE)	ZERO-SHOT ACCURACY EAMT22
Full MTCUE	**46.89**	**54.06**	62.67	**100.0**	81.35
no context encoder	46.76	53.73	**63.26**	89.10	77.42
no pos. embeddings	46.68	53.81	62.47	91.65	70.91
no MINILM-V2	45.32	53.42	62.55	50.00	70.16
no metadata	45.23	53.64	62.64	89.70	**83.41**
no doc.-level data	46.23	53.49	61.67	68.80	74.64
random context	42.17	51.94	61.74	49.90	68.44
no context*	41.22	50.07	58.94	50.00	67.53

Table 6: Ablation study on model components and data settings. *Corresponds to non-contextual Transformer.

We discuss the robustness of MTCUE with an ablation study on the model components as well as a complementary ablation on types of context (metadata vs document). We evaluate three language pairs (EN→DE,FR,PL) and report results from single runs (Table 6): COMET score on the OpenSub-

titles18 data and zero-shot accuracy at the two contextual tasks (on the **validation** sets in all cases).

Removing the context encoder (output of the linear layer is combined with source straight away) or the position embeddings has only a minor effect on the COMET score; replacing MINILM-V2 with a discrete embedding function hurts performance the most. Positional embeddings seem more important to the EAMT22 task than IWSLT22 - possibly because EAMT22 focuses on sentence-level phenomena, so the order of past context matters.

Replacing MINILM-V2 with a discrete embedding function removes the zero-shot effect in both tasks. An interesting finding is that between metadata and document-level data, it is the latter that brings more improvements to contextual tasks; this means that our model potentially scales to domains without metadata. Finally, using random context degrades performance w.r.t. full model implying that the gains come from signals in data rather than an increase in parameters or training time.

6 Related Work

Although contextual adaptation has been discussed in other tasks (e.g. Keskar et al., 2019), in this section we focus on NMT, as well as set our work side by side with research that inspired our approach.

Existing studies on incorporating context into NMT have primarily focused on document-level context. These approaches include multi-encoder models (e.g. Miculicich et al., 2018), cache models (Kuang et al., 2018), automatic post-editing (Voita et al., 2019a), shallow fusion with a document-level language model (Sugiyama and Yoshinaga, 2021), data engineering techniques (Lupo et al., 2022) or simple concatenation models (Tiedemann and Scherrer, 2017). Another line of research aims to restrict hypotheses based on certain pre-determined conditions, and this includes formality (Sennrich et al., 2016a), interlocutors' genders (e.g. Vanmassenhove et al., 2018; Moryossef et al., 2019), or a combination of both (Vincent et al., 2022b). Other conditions include translation length and monotonicity (Lakew et al., 2019; Schioppa et al., 2021), vocabulary usage (Post and Vilar, 2018) or domain and genre (Matusov et al., 2020). While wider contextual adaptation in NMT has been discussed theoretically, most empirical research falls back to gender (Rabinovich et al., 2017) or formality control (Niu et al., 2017). One exception is Michel and Neubig (2018b) who adapt NMT for each of many speakers by adding a "speaker bias" vector to the decoder outputs.

Our work is motivated by the CUE vectors (Novotney et al., 2022) and their application in personalised language models for film and TV dialogue Vincent et al. (2023). CUE vectors represent context computed by passing sentence embeddings of the input context through a dedicated encoder. Novotney et al. show that incorporating CUE in language modelling improves perplexity, while Vincent et al. use them to personalise language models for on-screen characters. In contrast, we reformulate CUE for contextual machine translation, provide a detailed analysis of incorporating CUE into the model, emphasise the importance of vectorising the context prior to embedding it, and examine the benefits for zero-shot and few-shot performance in contextual NMT tasks.

7 Conclusions

We have presented MTCUE, a new NMT architecture that enables zero- and few-shot control of contextual variables, leading to superior translation quality compared to strong baselines across multiple language pairs (English to others, cf. Table 4). We demonstrated that using sentence embedding-based vectorisation functions over discrete embeddings and leveraging a context encoder significantly enhances zero- and few-shot performance on contextual translation tasks. MTCUE outperforms the winning submission to the IWSLT22 formality control task for two language pairs, with zero-shot accuracies of 100.0 and 99.7 accuracy respectively, without relying on any data or modelling procedures for formality specifically. It also improves by 12.08 accuracy points over the non-contextual baseline in zero-shot control of interlocutor attributes in translation at the EAMT22 English-to-Polish task. Our ablation study and experiments on formality in English-to-German demonstrated that the context encoder is an integral part of our solution. The context embeddings produced by the context encoder of the trained MTCUE can be mapped to specific effects in translation outputs, partially explaining the model's improved translation quality. Our approach emphasises the potential of learning from diverse contexts to achieve desired effects in translation, as evidenced by successful improvements in formality and gender tasks using film metadata and document-level information in the dialogue domain.

Limitations

While we carried out our research in four language pairs (in both directions), we recognise that these are mainly European languages and each pair is from or into English. The choice of language pairs was limited by the data and evaluation tools we had access to, however as our methods are language-independent, this research could be expanded to other pairs in the future.

Another limitation is that the work was conducted in one domain (TV subtitles) and it remains for future work to investigate whether similar benefits can be achieved in other domains, though the findings within language modelling with CUE in Novotney et al. (2022) who used a different domain suggest so.

Ethics Statement

We do not foresee a direct use of our work in an unethical setting. However, as with all research using or relying on LMs, our work is also prone to the same unwanted biases that these models already contain (e.g. social biases). Therefore, when controlling contextual attributes, researchers should be aware of the biases in their data in order to understand the models' behaviour.

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of the High Performance Computing Service. This work was also supported by ZOO Digital.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv*.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. 2020. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv*, pages 1–18.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shaohui Kuang, 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. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Evgeny Matusov, Patrick Wilken, and Christian Herold. 2020. Flexible customization of a single neural machine translation system with multi-dimensional metadata inputs. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 204–216, Virtual. Association for Machine Translation in the Americas.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Paul Michel and Graham Neubig. 2018a. Extreme adaptation for personalized neural machine translation. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2:312–318.

Paul Michel and Graham Neubig. 2018b. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, 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, Brussels, Belgium. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. pages 49–54.

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

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Scott Novotney, Sreeparna Mukherjee, Zeeshan Ahmed, and Andreas Stolcke. 2022. CUE vectors: Modular training of language models conditioned on diverse contextual signals. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3368–3379, Dublin, Ireland. Association for Computational Linguistics.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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)*, pages 851–864, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for 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 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1:1074–1084.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling translation formality using pre-trained multilingual language models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 327–340, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 1:86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3:1715–1725.

Dagmar Stahlberg, F Braun, L Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5781–5791, Online. Association for Computational Linguistics.

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

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2016. Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2571–2581, Osaka, Japan. The COLING 2016 Organizing Committee.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3003–3008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5999–6009.

Sebastian Vincent, Loïc Barrault, and Carolina Scarton. 2022a. Controlling formality in low-resource NMT with domain adaptation and re-ranking: SLT-CDT-UoS at IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 341–350, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Sebastian Vincent, Rowanne Sumner, Alice Dowek, Charlotte Blundell, Emily Preston, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2023. Personalised language modelling of screen characters using rich metadata annotations. In *arXiv:2303.16618*. Preprint.

Sebastian T. Vincent, Loïc Barrault, and Carolina Scarton. 2022b. Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 121–130, Ghent, Belgium. European Association for Machine Translation.

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 877–886, 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.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021a. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Yue Wang, Cuong Hoang, and Marcello Federico. 2021b. Towards modeling the style of translators in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1193–1199, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Data Preprocessing

Parsing OpenSubtitles To prepare OpenSubtitles (specifically the document-level part of the corpus), we follow the setup described in Voita et al. (2019b). There are timestamps and overlap values for each source-target sample in the corpus; we only take into account pairs with overlap $>= 0.9$ and we use two criteria to build any continuous document: (1) no omitted pairs (due to poor overlap) and (2) no distance greater than seven seconds between any two consecutive pairs. To generate train/validation/test splits, we use generated lists of held-out IMDB IDs based on various published test sets (Müller et al., 2018; Lopes et al., 2020; Vincent et al., 2022b) to promote reproducibility. These lists can be found within the GitHub repository associated with this paper.

Embedding contexts Since a lot of metadata is repeated, and models are trained for multiple epochs, we opt for the most efficient way of embedding and storing data which is to use a memory-mapped binary file with embeddings for unique contexts, and an index which maps each sample to its embedding. This saves more than 90% space w.r.t. storing a matrix of all embeddings, and trains over $3\times$ faster than embedding batches on-the-fly.

B Model details

MTCue is trained from a pre-trained machine translation model (corresponding to the Base model) which is the `transformer` NMT architecture within Fairseq. We follow model specifications and training recommendations set out by Fairseq in their examples for training a translation model[8]. We train a model for each of the eight language directions on the source-target pairs from OpenSubtitles. We train the model until a patience parameter of 5 is exhausted on the validation loss.

C Observations on training and hyperparameters

We shortly describe here our findings from seeking the optimal architecture for MTCue and training settings in the hope that this helps save the time of researchers expanding on our work.

- Reducing the number of context encoder layers led to inferior performance.
- Freezing the source encoder when fine-tuning MTCue from a translation model led to inferior performance,
- Training MTCue from scratch – significantly increased training time while having a minor effect on performance.

[8] https://github.com/facebookresearch/fairseq/tree/main/examples/translation#iwslt14-german-to-english-transformer, accessed 1/5/23.

- Other context combination strategies (sequential and flat attention in Libovický et al., 2018) led to similar results.
- Some alternatives to QK-NORM to combat the problem of the exploding dot-product were successful but had a negative impact on performance:
  - using layer normalisation after the linear layer is applied to vectorised contexts,
  - using SmallEmb[9] which initialises the embedding layer (in our case, the linear proj. layer) to tiny numbers and adds layer normalisation on top.
- Zero-shot performance at the IWSLT22 task is generally consistent (at around $98.0-100.0$ accuracy) though may vary depending on the selected checkpoint. We found that training MTCUE for longer (i.e. more than 20 epochs) may improve translation quality but degrade the performance on e.g. this task.
- We found that MTCUE is generally robust to some hyperparameter manipulation on the OPENSUBTITLES dataset, and recommend performing a hyperparameter search when training the model on new data. For simplicity, in this paper we use a single set of hyperparameters for all language directions, though for some pairs the results may improve by manipulating parameters such as batch size and context dropout.

D Formality

To evaluate the performance of any tested model on the formality task we had to come up with a fair method of choosing a context to condition on, since in a zero-shot setting the model organically learns the tested attributes from various contexts rather than specific cherry-picked sentences.

To do so, we sampled some metadata from the validation set of the OPENSUBTITLES data and picked eight contexts (four for the *formal* case and four for the *informal* case) which either used formal or informal language themselves or represented a domain where such language would be used. We also added two generic prompts: *Formal conversation* and *Informal chit-chat*. The full list of prompts was as follows:

- Formal:
  1. *Formal conversation*
  2. *Hannah Larsen, meet Sonia Jimenez. One of my favourite nurses.*
  3. *In case anything goes down we need all the manpower alert, not comfortably numb.*
  4. *Biography, Drama,*
  5. *A musician travels a great distance to return an instrument to his elderly teacher*
- Informal:
  1. *Informal chit-chat*
  2. *I'm gay for Jamie.*
  3. *What else can a pathetic loser do?*
  4. *Drama, Family, Romance*
  5. *Animation, Adventure, Comedy*

We then ran the evaluation as normal with each context separately, and selected the highest returned score for each attribute.

E Examples of Model Outputs (Zero-Shot)

We include examples of translations produced zero-shot by MTCUE in Table 7. We would like to draw attention particularly to the top example for the EAMT22 task ("I just didn't want you to think you had to marry me"). The phrase *to marry someone* can be translated to Polish in several ways, indicating that the addressee is to be a wife (*ożenić się z kimś*), a husband (*wyjść za kogoś [za mąż]*) or neutral (*wziąć ślub*). While the reference in this example uses a neutral version, both the baseline model and MTCUE opted for feminine/masculine variants. However, the gender of the speaker is feminine, so the phrase "... *had to marry me*" should use either the neutral version (*wziąć ślub*) or the feminine one (*ożenić się*). The baseline model incorrectly picks the masculine version while MTCUE is able to pick the correct one based on the context given. MTCUE also correctly translates the gender of the interlocutor: both in the top example (*myślał* vs *myślała*) and the bottom one (*aś* vs *eś*, even though a synonymous expression is used in translation, agreement remains correct). Finally, the IWSLT22 example shows how MTCUE produces correct possessive adjectives for each formality.

[9] https://github.com/BlinkDL/SmallInitEmb, accessed 1/5/23.

	EAMT22
Source	I just didn't want you to think you had to marry me.
Context	*I am a woman. I am talking to a man*
Reference	Bo nie chciałam, żebyś myślał, że cię zmuszam do ślubu.
	"Because I didn't want$_{feminine}$ you to think$_{masculine}$ I am forcing you into a wedding."
Baseline	Po prostu nie chciałem, żebyś myślała, że musisz za mnie wyjść.
	"I just didn't want$_{masculine}$ you to think$_{feminine}$ you had to marry$_{feminine}$ me."
MTCUE	Nie chciałam, żebyś myślał, że musisz się ze mną ożenić.
	"I didn't want$_{feminine}$ you to think$_{masculine}$ you had to marry$_{masculine}$ me."
Source	So then you confronted Derek.
Context	*I am talking to a woman*
Reference	A więc doprowadziłaś do konfrontacji z Derekiem.
	"So then you led$_{feminine}$ to a confrontation with Derek."
Baseline	Więc wtedy skonfrontowałeś się z Derekiem.
	"So then you confronted$_{masculine}$ Derek."
MTCUE	Więc skonfrontowałaś się z Derekiem.
	"So then you confronted$_{feminine}$ Derek."
	IWSLT22
Source	I got a hundred colours in your city.
MTCUE (formal)	Ich habe 100 Farben in Ihrer Stadt.
MTCUE (informal)	Ich hab 100 Farben in deiner Stadt.

Table 7: Examples of MTCUE's outputs (zero-shot) versus a non-contextual Transformer baseline.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Unnumbered section after section 7 (Conclusions)

☒ A2. Did you discuss any potential risks of your work?
There are no relevant risks associated with our work

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract + section 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2 (proposed architecture); section 3.1 (used the OpenSubtitles corpus); section 3.2 (used two evaluation suites); section 2.2 (used sentence embedding models); section 3.4 (used software for implementation)

☑ B1. Did you cite the creators of artifacts you used?
Yes (sections 3.1, 3.2, 2.2, 3.4)

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
For created artifacts: section 1 For used artifacts: section 3.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 3.1. For sentence-transformers we did not explicitly discuss this but made all the necessary steps requested by the authors, such as citing the library and relevant papers.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 1, Table 2, sections 3.1, 3.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 3 (experimental setup), section 4 (results), section 5 (ablation study)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.4, Table 4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, section 3.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 2.2, 3.1, 3.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.