Reference-less Analysis of Context Specificity in Translation with Personalised Language Models

♣♦**Sebastian Vincent**, ♦**Alice Dowek**, ♦**Rowanne Sumner**, ♦**Charlotte Prescott**,
♦**Emily Preston**, ♦**Chris Bayliss**, ♦**Chris Oakley**, ♣**Carolina Scarton**

♣Department of Computer Science, University of Sheffield, UK

♦ZOO Digital Group PLC, UK

c.scarton@sheffield.ac.uk

Abstract

Sensitising language models (LMs) to external context helps them to more effectively capture the speaking patterns of individuals with specific characteristics or in particular environments. This work investigates to what extent detailed character and film annotations can be leveraged to personalise LMs in a scalable manner. We then explore the use of such models in evaluating context specificity in machine translation. We build LMs which leverage rich contextual information to reduce perplexity by up to 6.5% compared to a non-contextual model, and generalise well to a scenario with no speaker-specific data, relying on combinations of demographic characteristics expressed via metadata. Our findings are consistent across two corpora, one of which (CORNELL-RICH) is also a contribution of this paper. We then use our personalised LMs to measure the co-occurrence of extra-textual context and translation hypotheses in a machine translation setting. Our results suggest that the degree to which professional translations in our domain are context-specific can be preserved to a better extent by a contextual machine translation model than a non-contextual model, which is also reflected in the contextual model's superior reference-based scores.

1. Introduction

Contrary to studies in sociolinguistics (Milburn, 2004), conventional approaches to generation tasks build models in a one-size-fits-all fashion, and most often for a particular language and domain, disregarding the context of the processed text. This leads to assuming the most likely scenario as context, sometimes resulting in harmful predictions (e.g. the "masculine default" in Schiebinger, 2014). **Personalisation** – adapting model predictions to the unique dialogues of individuals – offers clear benefits in generation tasks (Flek, 2020), where context information helps to disambiguate the input text, aiding correct interpretation and minimising sample bias in training data (Dudy et al., 2021).

Incorporating demographic factors has been shown to improve the performance of models developed for NLP tasks such as classification (Hovy, 2015) and translation (Vincent et al., 2022b). Their influence can manifest in two main dimensions: **grammatical** and **behavioural**. Grammatical agreement is well-defined and pertains to morphosyntactic structures, whereas behavioural is more fluid, relating to how language is utilised across different demographics or contexts (Vincent, 2023). For instance, the speaker's gender may dictate morphological endings in self-referential verbs in certain languages, exemplifying grammatical agreement. Conversely, behavioural agreement is more nuanced: "They're done!" has a different meaning when said by a baker about a batch of cookies than by a frenzied king about his treacherous subjects. Contextual language generation research (e.g. Voita et al., 2019) mostly emphasises grammatical agreement, whereas in practice both types of agreement are required in the language generation process and a robust framework should encompass them both.

This work explores personalisation of language models (LMs) and machine translation (MT) systems for **speakers** and **productions** in TV series and film, focusing on leveraging the similarities in their metadata, an approach termed **context-based** personalisation. The language usage within our domain exhibits significant variability. For instance, TV writers often create characters that emulate the speech patterns of particular social groups they represent. Similarly, productions from specific decades, countries, or genres capture unique discourse nuances. We demonstrate how speaker and production metadata can inform the development of LMs which excel at capturing the language style of a given speaker or production compared to generic models. We then apply these LMs to measure the **context specificity** of hypothesis tokens used in professional and machine translations within the TV series domain. This evaluation is done in a reference-less manner, taking the step towards a much-needed evaluation tool for context-aware MT free of reference bias.

Our domain presents an additional challenge: ensuring model robustness in scenarios where no prior dialogue samples exist for certain speakers or productions. This setting, where there is insuffi-

cient content to characterise the subjects of a given system, is commonly known as the cold start problem (e.g. Schein et al., 2002; Huang et al., 2014). Models adapted solely on past dialogue fall short in this case, and we argue that a context-based approach is more effective, mimicking the benefits of personalisation by estimating token distributions for similar character/production profiles.

We collect a rich set of metadata annotations for two corpora based on identifiable information (i.e. character names and titles). This allows us to perform experiments on up to 14 unique metadata variables at once, to our knowledge the richest set of metadata information for personalisation. In contrast, metadata-based approaches to personalisation reported in previous work in different domains were small-scale, leveraging a few simple and mostly categorical variables (Huang et al., 2014; Lynn et al., 2017; King and Cook, 2020; Welch et al., 2020; Guo et al., 2021).

Our work is presented in two parts: firstly, we consider whether detailed character profiles can be used to model the characters' speaking styles; we include a case for characters excluded from the training data, for whom the models learn adaptation based on characters with similar dossiers (§3). Then, we explore how such personalised LMs can be used to estimate the context specificity (or extent of personalisation) of professional and machine translations (§4). Additionally, we contribute CORNELL-RICH (§3.1), a corpus of rich character and film annotations for the Cornell Movie Dialogue Corpus (Danescu-Niculescu-Mizil and Lee, 2011) (CORNELL) and SMRR, an evaluation metric for personalised LMs.[1] This paper also presents the related work (§2), and conclusions (§5).

2. Related Work

Personalisation in NLP Personalisation in NLP can generally be split into three groups: with respect to how much data is available for a speaker: *full supervision*, where there is sufficient training data to fine-tune a model for a particular speaker, *few-shot*, where some supervised data exists but not in quantities sufficient for supervised training, and *zero-shot*, where no samples of text exist for the speaker used in evaluation. Full supervision is usually facilitated through some form of a *user embedding* or *tagging* approach (e.g. Sennrich et al., 2016; Keskar et al., 2019; Mireshghallah et al., 2022). Among few-shot approaches, King and Cook (2020) examine several personalisation methods for language modelling of blog posts with sample adaptation data for new users, including fine-tuning, interpolation (averaging the

fine-tuned speaker model with a general model) and demographic-based adaptation (fine-tuning on text from users with the same age and gender). Welch et al. (2022) leverage models built for "anchor users" (who boast a large history of posting) to devise models for new users (with a small number of posts), focusing on the similarity between samples of users' posts. Finally, zero-shot approaches typically leverage background data available for the new speakers, e.g. their metadata. Huang et al. (2014) rely on the social network of a user to model their language; Lynn et al. (2017) use age, gender, and personality traits to improve user modelling in multiple NLP tasks; Zeng et al. (2019) leverage user profiles to improve comment generation on a social media corpus. Demographic-based adaptation was found inferior to interpolation and priming in the few-shot scenario by King and Cook (2020), but their study used only two factors, age and gender. Our work is positioned in the zero-shot category as we rely on rich metadata annotations to model the dialogue of individual screen characters appearing in particular productions. Unlike King and Cook (2020), we leverage textual (real-valued) metadata annotations, which in personalisation are preferable to categorical values (Lynn et al., 2017), and a significantly higher count of them (up to 14). Importantly, we also leverage our personalised LMs to quantify context specificity in translation.

A few studies have explored the idea that context in NLP can be summarised with pre-trained models. Novotney et al. (2022) introduce the notion of CUE (contextual universal embeddings), representing individual context variables as pre-trained sentence embedding vectors. They demonstrate that including article metadata in the form of CUE into language modelling improves perplexity when applied to news articles. Vincent et al. (2023) explore this idea further, applying it to machine translation of dialogue and showing that pre-training on film metadata helps zero- and few-shot performance in some contextual MT tasks. In this paper, we leverage context in the same way as Vincent et al. (2023), but focus on contextual language modelling, and specifically on personalisation for individual characters and films. We also explore a practical application of such personalised LMs in evaluation of contextual MT (§4), and contribute an evaluation metric for personalised LMs focused on per-speaker performance (§3.2).

Evaluation of Contextual Machine Translation Traditional measures of MT quality are based on sentence-level matching to references, and offer little insight into performance at maintaining or introducing context-specific features of the source text. Alternative evaluation methods of contextual MT have been explored to address this. When

---

contextual phenomena are directly observable and necessitate grammatical agreement (e.g. in formality transfer or document-level translation), evaluation usually involves parsing tools (Sennrich et al., 2016; Vincent et al., 2022a) or contrastive evaluation on bespoke test suites (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019; Lopes et al., 2020). However, the creation of such tools and test sets is expensive, and as argued in Post and Junczys-Dowmunt (2023), strong performance at contrastive evaluation does not necessarily entail the ability to generate contextual translations in practice. Evaluation of behavioural agreement (e.g. preserving individual style of a character or production), has mostly been limited to classification systems (e.g. Michel and Neubig, 2018) which attribute the input text as belonging to one of a list of speakers. However, such systems depend on sufficient quantities of training data for each considered speaker, which is usually not readily available.

Language Models in Machine Translation LMs have been utilised in machine translation as means of improving fluency (Stahlberg et al., 2018), boosting document-level performance (e.g. Sugiyama and Yoshinaga, 2021) or evaluation (Edunov et al., 2020). In contrast, we build a tandem of LMs from the same training data as the NMT models to measure to what extent the NMT generations are context-specific. Our metric is pointwise mutual information (PMI), similarly to Sugiyama and Yoshinaga (2021) who used PMI between document context and the target utterance to boost document-level performance of MT.

3. Building a Personalised LM

The first stage of our work delves into building a personalised language model for dialogue associated with rich contextual annotations. We create two metadata-rich datasets (§3.1) and train a contextual language model to capture the distribution of the tokens in the dialogue given a set of contextual variables (§3.2). This section addresses two research questions (RQs):

RQA *How can rich character profiles be used to model the characters' speaking styles?*

RQB *How can a LM be personalised for a specific character solely by learning from data for characters with similar profiles?*

3.1. Datasets

In the absence of similar datasets, we use a combination of manual and automatic annotation to create two English-language corpora: ZOO and CORNELL-RICH. The domain of both corpora is

TV series and film dialogue respectively, and samples within each corpus consist of: an utterance in English and a set of up to 14 textual metadata annotations for the speaking character (age bracket, country of origin, description, gender, profession, religion and characteristic quote) and for the production (country, genre, PG rating, plot description, writers, year). All metadata contained in the corpora was collected by the authors of this paper manually or automatically. Metadata within CORNELL-RICH (the corpus we make publicly available) was collected from readily available materials with a permissive license. Below we summarise the descriptions for each corpus.

ZOO The ZOO corpus is a private in-house collection of subtitles for nine anglophone TV series. It totals 157K dialogue lines and annotations for 159 speakers of 101K lines. It is divided into traditional `test`, `valid` and `train` sets, but features an additional test set of metadata and dialogue from 11 held-out speakers who do not appear in the remaining sets. Quantitative details are reported in Table 1 (rows 1-5). The corpus was created from production-ready subtitle files from which dialogue with character and TV series attributions was extracted. This data was subsequently annotated with production metadata (automatically, via the OMDb API[2]) and character metadata for the most frequently speaking characters. The annotation process is detailed in Appendix A. In total, 87.1% of productions and 50.3% of speakers received annotations.

CORNELL-RICH Much like ZOO, CORNELL-RICH is a dataset of rich character and production annotations, albeit for film dialogue extracted from scripts. It includes 14 distinct metadata variables captured as text. The collected annotations can be linked to the entries of CORNELL (Danescu-Niculescu-Mizil and Lee, 2011), a corpus of exchanges from a set of film scripts, with character dialogue attributions (Figure 1 illustrates how CORNELL-RICH enriches the original corpus). Both dialogue data and annotations are in English. We provide details about the collected metadata, together with examples, in Appendix B. The corpus is split into `train`, `valid`, `test` and `test_unseen` in the same way as ZOO, and `test_unseen` features 30 held-out speakers (see Table 1: rows 6-10).

OPENSUBTITLES We use the OpenSubtitles18[3] corpus (Lison et al., 2018) (OPENSUBTITLES) to pre-train the LMs. It is a a large collection of subtitles with timestamps that facilitate the extraction
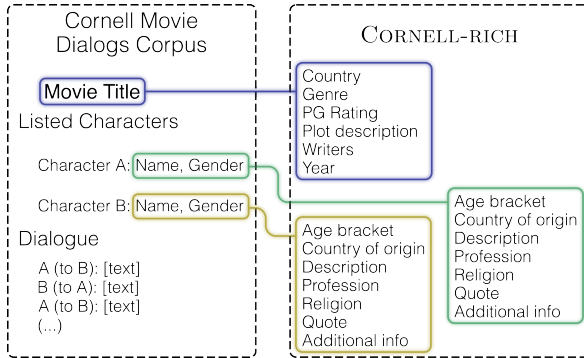

Figure 1: CORNELL-RICH overview.

of document-level information. Focusing on past context with no loss of generality, we extract up to 3 past sentences based on the timestamps (Table 1, rows 11-12). Roughly 68% samples contain at least one past sentence.

Preprocessing Since both ZOO and CORNELL are of high quality as is, our preprocessing only involves normalising punctuation, removing tokenisation using the `sacremoses` package[4], fixing leftover punctuation issues (e.g. ensuring all multi-dots use three dots) and removing HTML tags. We also preprocess all (original and added) annotations so that: (i) all empty fields are expressed as an empty string; (ii) there are no multiple expressions of the same discrete type (e.g. *m* and *M* to denote masculine gender); (iii) all attributes are expressed in unambiguous natural language (e.g. a PG rating of "R" is rewritten as "PG Rating: R"). OPENSUBTITLES is preprocessed following Vincent et al. (2023). For subword tokenisation, we use SentencePiece to train a BPE model of 8K tokens on the `train` split of CORNELL-RICH; it is then used to tokenise all datasets.

| Row | Dataset & split | Total number of | | |
		segments	tokens	metadata types
(1)	ZOO			
(2)	train	140.4K	1.1M	
(3)	valid	4K	31.3K	
(4)	test	6K	47.1K	13
(5)	test_unseen	6.7K	51.5K	
(6)	CORNELL-RICH			
(7)	train	289.0K	3.1M	
(8)	valid	5K	51.2K	
(9)	test	5K	54.4K	14
(10)	test_unseen	5.2K	54.6K	
(11)	OPENSUBTITLES			
(12)	train	14.7M	109.6M	3*

Table 1: Quantities of segments, tokens (pre-tokenisation) and unique metadata (speaker and production) in datasets. *OPENSUBTITLES uses three past sentences as proxy metadata.

---

[4] https://pypi.org/project/mosestokenizer/

3.2. Experimental Setup

LMCUE Architecture Our selected LM architecture is adapted from the MTCUE model (Vincent et al., 2023), which is a Transformer-based multi-encoder contextual machine translation system. MTCUE processes the source text with a source encoder and the context information with an additional context encoder. We convert MTCUE to a LM by removing the source encoder, resulting in a conditional encoder-decoder LM where context is treated as the input to the encoder. The sequence of context information is converted to a sequence of equal-sized vectors with a sentence embedding model (MINILM-V2). This approach has the advantage of treating both discrete and continuous (text) inputs in the same way, potentially utilising the semantic information of the discrete labels, as well as allowing longer spans of context as input without issues of long-range dependencies. The target sequences are contextualised via standard encoder-decoder attention which maps queries (target) to keys and values (context). We select this approach as MTCUE can process large sets of contextual information and has the potential to scale well to few- and zero-shot scenarios, which in our case are explored when we consider test sets with completely new speakers. For simplicity, hereinafter we refer to this architecture as **LMCUE**.

Pre-training Preliminary experiments showed that training LMCUE from scratch on CORNELL-RICH lead to results inferior to a non-contextual LM trained on the same data (see Appendix C). We, therefore, experimented with pre-training the model first. Since a larger corpus of dialogue with character metadata is unavailable, we used a corpus with document-level information and treated the **past dialogue** for any sentence as the **metadata context**. We hypothesised that at a larger scale, the effect of metadata embeddings on text generation will be similar to the effect of embeddings of past dialogue (Figure 2), meaning the pre-training procedure allows the model to learn dependencies between the context and the text. An advantage of this approach is that metadata-rich corpora are likely to be too small to train a model from scratch, but document-level information is abundant. In our case, pre-training on past dialogue proved successful; consequently, all models considered in our experiments are pre-trained. For this purpose, we use the OPENSUBTITLES corpus (§3.1).

Baselines and Implementation We consider three baselines: a non-contextual LM (BASE-LM), a speaker-wise fine-tuning baseline (SPFINETUNING) and a linear interpolation method (LERP) which ensembles SPFINETUNING with the
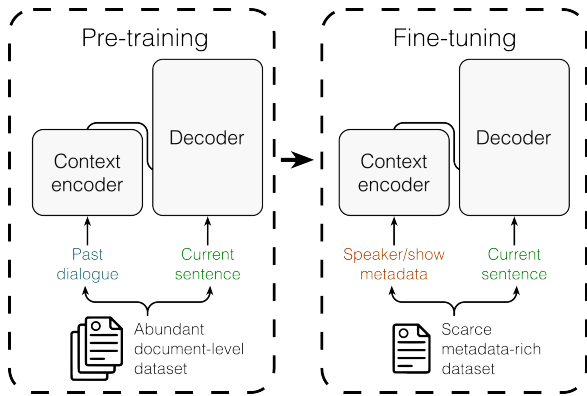
Figure 2: An illustration of the pre-training and fine-tuning regimens used in the experiments.

general model BASE-LM at test time (baselines are modelled after King and Cook, 2020).

We implement LMCUE by modifying the code provided by (Vincent et al., 2023). The model has 159M parameters and comprises a context encoder (38M) and a decoder (121M). 25% of the decoder's parameters are used by the encoder-decoder attention; a non-contextual decoder of this shape would have 91M parameters. To make the comparison fair, BASE-LM matches the total number of parameters in LMCUE (159M) and is therefore wider than the decoder in LMCUE (Table 2). This strong baseline removes the possibility that the model improves simply because of a higher parameter count. All other baselines (SPFINETUNE, LERP) share the architecture and size of BASE-LM.

		Params	d_{model}	n_{layers}	h	FFN dim.
(1)	LMCUE (Enc.)	38M	512	6	8	2048
(2)	LMCUE (Dec.)	121M	768	12	12	3072
(3)	LMCUE (total)	159M	–	–	–	–
(4)	BASE-LM	159M	1024	12	16	4096

Table 2: Model details for LMCUE and BASE-LM.

The LMCUE models are pre-trained on OPEN-SUBTITLES (using past dialogue as context). BASE-LM is pre-trained on the text part of the same corpus, one sentence at a time. Training details are provided in Appendix D.

Evaluation For evaluation, we use perplexity (PPL) and speaker mean reciprocal rank (sMRR), which we define as follows: let M_j be a model personalised for a speaker s_j and U_i be a set of utterances by a speaker s_i. We calculate speaker reciprocal rank sRR for any speaker k by scoring the U_k with $M_1, ..., M_n$ (expressed with log likelihood), then ranking the models best to worst by this score, with pessimistic tie resolution, and taking the reciprocal rank ($1/rank$) of M_k, the model for speaker k (see Figure 3). sMRR is sRR av-

eraged for all speakers; $1/$sMRR is the average rank of the correct speaker model. Intuitively, this metric captures the strength of the association between dialogue and the speaker model: sMRR of 1.0 indicates that for any speaker j, the model M_j produces the best score for U_j.
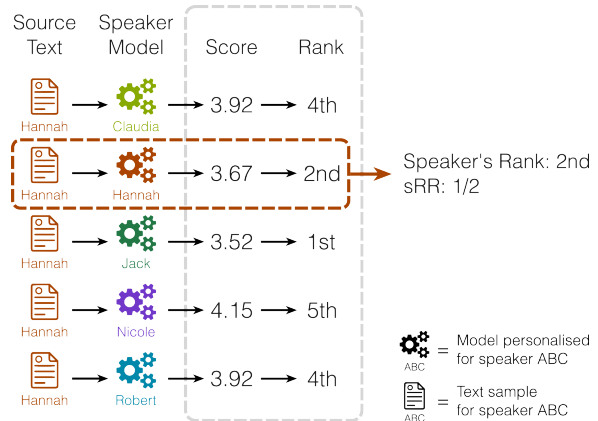

Figure 3: sRR illustrated for speaker Hannah.

Unless otherwise specified, all results are calculated from five runs with different random initialisation, and the reported value is the **mean** result. We **highlight** the best overall result. Unless another result is underlined, it is significantly worse (indicating a less effective model) than the best result in bold, with a confidence interval of 95% (computed with a one-tailed t-test, $t(4) = 1.65$, $p = 0.05$).

3.3. Results

This section presents the results of training LMCUE on the CORNELL-RICH and ZOO corpora. We use $\mathcal{S}$ and $\mathcal{P}$ to denote that $\mathcal{S}$peaker or $\mathcal{P}$roduction metadata was used in training (or both, i.e. $\mathcal{S} + \mathcal{P}$).

Are speaker profiles helpful? We examine whether including speaker profiles as a supplementary input in language modelling can result in significant quantitative improvements. For this, we train models on the `train` splits and evaluate on the `test` splits of both corpora, with overlapping speakers (= unique combinations of speaker profiles) between them. As demonstrated in Table 3, context-based personalisation with LMCUE results in substantial reductions in perplexity compared to the best baseline, with a decrease of 5.4% for CORNELL-RICH and 6.5% for ZOO, respectively.

Is speaker-based adaptation better than direct fine-tuning? To determine the effectiveness of our context-based adaptation approach when compared to LMs fine-tuned on the available speaker dialogue, we focus our setup on five long-term (multi-episode TV) speakers with at least 3,000

	CORNELL-RICH		ZOO	
	valid	test	valid	test
baselines				
BASE-LM	22.35	23.38	18.42	18.41
proposed				
LMCUE ($\mathcal{S}$)	21.37	22.37	17.52	17.55
LMCUE ($\mathcal{P}$)	**21.07**	**22.04**	17.18	17.29
LMCUE ($\mathcal{S} + \mathcal{P}$)	21.14	22.13	**17.13**	**17.21**

Table 3: Perplexity$^{\downarrow}$ on different validation and testing sets for the two corpora.

lines of dialogue sources from the ZOO corpus (Table 4). For each speaker, we use 400 and 600 of these lines for validation and testing, respectively. Within this experiment, we use SPFINETUNE and LERP as baselines. We obtain SPFINETUNE by fine-tuning the LM on all ZOO data initially (FT1), and then once more on speaker data alone. LERP is obtained from a mean interpolation of the prediction probabilities from SPFINETUNE and FT1. We fine-tune an individual model instance for each speaker $\in \{$**sp01**, ..., **sp05**$\}$.

LMCUE achieves results comparable to all speaker-fine-tuned models (Table 6). When using speaker metadata ($\mathcal{S}$), LMCUE achieves sMRR of 1.0 just like fine-tuned models, suggesting the perplexity improvements come from the model's context-based predictions. LMCUE ($\mathcal{P}$) achieves lower sMRR (0.8): its predictions are based only on production metadata, not considering that two different characters may come from the same production. Any adapted model, whether fine-tuned or metadata-based, yields a reduction in perplexity between 5.1% and 6.8% which is comparable to results on test. SPFINETUNE achieves the best overall perplexity reduction of 1.32 and 1.0 sMRR, with LMCUE ($\mathcal{S} + \mathcal{P}$) yielding a statistically comparable reduction of 1.29 and the same sMRR while requiring (i) no fine-tuning and (ii) the maintenance of only one model for all speakers.

To illustrate how personalisation manifests in practice, we identify the predictions of LMCUE ($\mathcal{S} + \mathcal{P}$) with the most increased and decreased log likelihood compared to BASE-LM (compare Table 4 and Table 5). Top-gaining tokens have strong associations with certain categories, like *profession* (**sp01** "Student, Spy" $\rightarrow$ *spy, Mom, mission*), *age* (**sp02** "Young Adult" $\rightarrow$ expletives) or *genre* (**sp04, sp05** "Crime, Drama, Horror" $\rightarrow$ *killer*). Similarly, top-gaining sentences for **sp01** and **sp02** have a comedic overtone (matching the genre), while the top-losing sentences do not fit these characters' demographic profiles.

For further analysis on this subject, Appendix E shows the individual contribution of each character metadata type to reduction in perplexity.

Zero-shot Transfer We assess the effectiveness of speaker adaptation for completely **new test speakers** featured in the test_unseen sets of both corpora. To reiterate, these speakers' dialogue is excluded from training and validation data (although there are overlaps in production metadata). As before, we fine-tune the pre-trained LMCUE on the train splits. We compare the performance only to BASE-LM since other baselines are not equipped to work well in this zero-shot scenario.

Table 7 shows that LMCUE ($\mathcal{S}$) still improves perplexity over a parameter-matched LM. Though these improvements are smaller than in the supervised scenario, they are still significant, especially for ZOO (-0.62). More importantly, for both corpora $\mathcal{S}$ is strongly beneficial towards high speaker separation (i.e. the model assigns the highest probability to dialogue which matches the given speaker's profile), as measured by sMRR. Perplexity does improve more when $\mathcal{P}$ is also used ($1.4 \rightarrow 5.6\%$ for CORNELL-RICH, $3.6 \rightarrow 4.4\%$ for ZOO), though in this scenario we are evaluating the easier task of modelling new speakers in seen or unseen productions. Production metadata alone yields the best reduction of 6.9/5.3%. Using it has a different effect on the two test sets: in CORNELL-RICH it induces a stronger boost in sMRR than $\mathcal{S}$ ($+0.08$), while in ZOO it decreases it considerably (-0.20); second, using it in conjunction with $\mathcal{S}$ results in best sMRR in CORNELL (0.94), but not so for ZOO. This can be explained by the fact that ZOO uses a pool of only nine productions (vs 595 in CORNELL-RICH), so adding $\mathcal{P}$ on top of $\mathcal{S}$ is unlikely to increase speaker separation. In contrast, CORNELL-RICH uses a rich pool of films, so film metadata is more likely to be unique between any two speakers, thus introducing it separates the two speakers even more, increasing sMRR. This is also why sMRR is so high for LMCUE ($\mathcal{P}$): with 24 unique films between the 30 speakers the film metadata is rarely shared between any two speakers, making their context inputs more dissimilar. The magnitude of improvements in sMRR is also different for the two corpora, which again could be attributed to scale (863 vs 159 speakers, 595 vs 9 productions). Increasing the number of annotated entities can therefore improve the personalisation effect. Nevertheless, a score of 0.39 still suggests that LMCUE ranks an unseen character on the 2.56th position with a model built from their demographic profile, on average.

Using LMCUE ($\mathcal{S} + \mathcal{P}$), we queried the words for which log likelihood increased the most w.r.t. BASE-LM in the test set of CORNELL-RICH and obtained a list of the following fifteen tokens:

crew shark ship azz birds casino space leads power ocean camp boat cops baby ace

Many of these tokens are context-specific and

ID	#Lines	Age	Profession	Country	Genre	Characteristic quote
sp01	7.5K	Teen	Student, Spy	United States	Comedy	"Look, I love you! I have loved you since the moment I saw you. Please! I'll never get a chance to tell you."
sp02	3.9K	Young Adult	Unemployed, Community Service	United Kingdom	Comedy, Drama	"In the words of the great Lionel Richie...hello."
sp03	3.2K	Adult	Actor	United States	Docuseries	"So be present, be spontaneous. Enjoy the moment, enjoy yourself and learn."
sp04	3.1K	Adult	Criminal Profiler	United States	Crime, Drama, Horror	"It isn't very smart to piss off a guy who thinks about killing people for a living."
sp05	3.1K	Adult	Psychiatrist	United States	Crime, Drama, Horror	"Before we begin, I must warn you... nothing here is vegetarian."

Table 4: Selected metadata regarding long-term speakers from ZOO used in the experiment.

ID	Top-gaining sentence (4+ words)	Five top-gaining tokens	Top-losing sentence (4+ words)
sp01	"Paranoid and can fit into small spaces."	Okay, Wait, spy, Mom, mission	"To teach and to lend a guiding hand."
sp02	"Fucking nuns! Fucking shit!"	Fuck, Shit, Fucking, fucking, fuck	"English, Math and French."
sp03	"I love this car."	Wow, ital, coffee, brain, b	"I'm not opposed to doing things to my teeth."
sp04	"One missing kid's a boy."	killer, Jack, kill, close, life	"She was a slim and delicate pig."
sp05	"Is your conscience clear?"	got, killer, Will, Jack, Ab	"Simpler times in boatyards with dad."

Table 5: Sentences and tokens for which the log likelihood under LMCUE $(\mathcal{S} + \mathcal{P})$ changes the most compared BASE-LM.

	sMRR$^\uparrow$	PPL$^\downarrow$					
		sp01	sp02	sp03	sp04	sp05	Mean
baselines							
non-context							
BASE-LM	0.2	15.24	17.39	23.53	18.64	21.14	19.19
fine-tuning							
SPFINETUNE	1.0	14.54	**16.01**	21.76	17.36	<u>19.50</u>	**17.84**
LERP	1.0	**14.35**	16.31	22.25	17.66	19.91	18.10
proposed							
metadata-based							
LMCUE $(\mathcal{S})$	1.0	14.99	16.75	21.86	17.54	19.89	18.21
LMCUE $(\mathcal{P})$	0.8	14.68	17.17	<u>21.26</u>	<u>17.12</u>	**19.45**	17.94
LMCUE $(\mathcal{S} + \mathcal{P})$	1.0	14.77	16.77	**21.22**	17.10	<u>19.47</u>	<u>17.87</u>

Table 6: Results on the test set for long-term speakers. <u>Underlined</u> results are on par with results in **bold**.

	CORNELL-RICH		ZOO	
	test_unseen		test_unseen	
Approach	PPL$^\downarrow$	sMRR$^\uparrow$	PPL$^\downarrow$	sMRR$^\uparrow$
baselines				
BASE-LM	23.62	0.03	17.11	0.09
proposed				
LMCUE $(\mathcal{S})$	23.28	0.70	16.49	**0.39**
LMCUE $(\mathcal{P})$	**22.00**	0.80	**16.21**	0.19
LMCUE $(\mathcal{S} + \mathcal{P})$	22.31	**0.96**	16.35	0.32

Table 7: Results of evaluation with speaker & film metadata on the test set of unseen speakers.

would only appear in certain scenarios or domains. For example, *casino* or *space* are unlikely to appear in a sentence unless they represent locations within the film. A subset of the provided tokens (*crew, shark, ship, ocean, birds*) may also collectively describe a single scenario, such as an adventure or thriller film set on a ship in the middle of an ocean. We hypothesise that a few such films appeared in the training set CORNELL-RICH, allowing LMCUE to develop a strong prior for predicting these tokens when metadata of similar films is provided as input. Finally, these tokens are notably more generic than those in Table 6: we observe that the effect of biasing speaker-specific vocabulary may be limited for some tokens compared to the supervised scenario (e.g. tokens representing names of the character's co-stars are not related to demographic features so would not be affected in a zero-shot scenario).

4. Measuring Personalisation in Translation

In §3 we have established empirically that LMCUE exhibits effects of context-based personalisation, acting as a person- and production-specific LM when provided with their metadata, and is comparable with speaker-specific fine-tuning approaches. Compared to a general LM, it assigns higher probability to tokens which are more likely to occur in the given character and production context. Within this section, we use this model as a "contextual oracle", applying it to various streams of dialogue to obtain judgements on how likely the dialogue is to be said in the given context. We also use a non-contextual LM as a "non-contextual oracle", to measure the extent to which the given text co-occurs specifically with the provided context. We are interested in the following research question:

RQC *Can MT offer personalisation benefits proportional to professional translations?*

We operate on four iterations of the same text: original version in English (ORIGINAL), professional translations of the original text to French, German or Polish (REFERENCE), and several versions of machine-translated text, which we describe below. Our goal is to establish to what extent the effect of personalisation (context-specificity to particular character and production descriptions) is found in professional and machine translations, and whether hypotheses generated by a contextual machine translation system exhibit stronger personalisation effects compared to non-contextual.

4.1. Datasets

The ZOO-MULTI Corpus We perform the experiments on ZOO-MULTI, which is a subset of the episodes featured in the ZOO corpus, but which includes professional translations to French, German and Polish. This corpus is split differently into training, validation and testing subsets than ZOO (Table 8). Specifically, we pre-select three TV series to maximise the variety of studied content within the domain: an unscripted documentary, an unscripted cooking competition show and a scripted drama series and use between four and six episodes each for testing. Then we split the remaining data as follows, in two different data settings: DISJOINT, where we use the remaining episodes as validation data, and all other TV series as training data (so that there is no overlap between training and validation/testing data) and OVERLAP, where we include the remaining episodes in the training data, and select random utterances from other series for validation. These two data settings represent the real-life scenarios of the subtitle translation task: when no past episodes are available for the considered series (DISJOINT), and when there are some already completed translations that can be leveraged (OVERLAP) – see Table 8, rows 1-8 for quantitative details and Appendix F for an individual sample.

OPENSUBTITLES Corpus For pre-training, we re-use the version of OPENSUBTITLES corpus described in Vincent et al. (2023). The dataset comprises sentence pairs annotated with six production metadata (via OMDb) and document-level data. Data quantities are listed in Table 8, row 10.

4.2. Evaluation and Systems

We evaluate how well adapted to context individual version of the text are by finding out the degree of co-occurrence between individual sentences or translations and their specific extra-textual context.

		Number of samples		
Row	Dataset & split	EN-FR	EN-DE	EN-PL
(1)	ZOO-MULTI			
(2)	DISJOINT			
(3)	train*	58.5K	59.0K	107.1K
(4)	valid*	4.0K	3.8K	4.1K
(5)	OVERLAP			
(6)	train*	60.3K	60.8K	106.1K
(7)	valid*	2.3K	2.3K	2.3K
(8)	test	7.8K	7.8K	7.6K
(9)	OPENSUBTITLES			
(10)	train	14.7M	5.3M	12.4M

Table 8: Quantities of segments in ZOO and OPEN-SUBTITLES. *Values are averaged over dataset iterations generated for each of the three series.

We express results as the pointwise mutual information (PMI) between the context $\mathcal{C}$ and the target utterance or hypothesis $\mathcal{H}$, which is computed as:

$$\text{PMI}(\mathcal{C}, \mathcal{H}) = \log \frac{p(\mathcal{H} \mid \mathcal{C})}{p(\mathcal{H})} \quad (1)$$

$$= \log p(\mathcal{H} \mid \mathcal{C}) - \log p(\mathcal{H}) \quad (2)$$

PMI rewards positively those tokens which occur more frequently in the context $\mathcal{C}$ than in the general distribution termed with the prior $p(\mathcal{H})$. In practice, both terms of Equation 2 are computed with LMs: $\log p(\mathcal{H} \mid \mathcal{C})$ with LMCUE $(\mathcal{S} + \mathcal{P})$ and $\log p(\mathcal{H})$ with BASE-LM. Both LMs are pre-trained as described in §3.2 and fine-tuned on the context and target-side dialogue from ZOO-MULTI corpus (OVERLAP setting). We train a separate tandem of LMs for each language pair, and for statistical significance we train five distinct instances of each model (each with a different random seed).

We use the open source implementation of MTCUE (Vincent et al., 2023) to generate non-contextual and contextual machine translations. We also follow the authors' implementation of a parameter-matched baseline (BASE-NMT). We pre-train the systems on OPENSUBTITLES with metadata and document-level information (as per the original paper) and fine-tune them on the ZOO-MULTI corpus for each language pair separately. We include training details in Appendix D.

4.3. Results

A positive value of PMI for REFERENCE (mean score of 0.073; Table 9) suggests the presence of a co-occurrence effect between professional translations of the test set and their extra-textual context. We calculated that the ORIGINAL text in English obtained a score of 0.087. The two values are not directly comparable (as computed on different language variants of the test set), but they are of

similar magnitude which suggests that the context-specific traits of the original text are well preserved in the professional translations for this test set.

	French	German	Polish	Mean
REFERENCE	0.101	0.037	0.081	0.073
DISJOINT				
BASE-NMT	0.042	−0.004	0.041	0.026
MTCUE	0.066	0.007	0.049	0.041
OVERLAP				
BASE-NMT	0.040	−0.006	0.049	0.028
MTCUE	**0.069**	**0.012**	**0.072**	**0.051**

Table 9: PMI computed with general and personalised LMs on translations. Results computed from five different runs.

	French		German		Polish	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
DISJOINT						
BASE-NMT	34.89	23.69	35.71	28.59	31.13	31.50
MTCUE	35.73	25.81	36.22	29.29	31.62	32.66
OVERLAP						
BASE-NMT	35.06	23.75	36.15	29.75	31.59	**32.83**
MTCUE	**36.14**	**27.04**	**36.90**	**30.53**	**32.18**	31.95

Table 10: BLEU and COMET scores for the evaluated MT systems. COMET score was computed using the `lrec20-comet-da` model.

BASE-NMT achieves a significantly lower PMI on average (Table 9, 0.026 for DISJOINT, 0.028 for OVERLAP): the absence of context at generation time results in translations less adapted to the specific characters and productions. The nevertheless positive values can be explained by the presence of domain-specific terms such as "yoghurt" which do not need context to be translated correctly but will occur more often in specific contexts (e.g. cooking shows), yielding a positive PMI. The contextual MTCUE achieves higher PMI on average than BASE-NMT (+0.015 for DISJOINT, +0.023 for OVERLAP), meaning that using the relevant context does make the hypotheses more personalised, and the greater improvement for OVERLAP suggests that context can be utilised even better when previous samples from the same speakers or series are given. Interestingly, while personalisation is stronger in OVERLAP systems than DISJOINT, the MTCUE (DISJOINT) still performs better than either BASE-NMT system, signifying the robustness of CUE vectors to this zero-shot adaptation to new series and speakers. Among all three target languages, reference translations to German are correlated with context the least (PMI = 0.037) and translations to French the most (PMI = 0.101). The MT systems' results follow a similar trend. We also provide the BLEU and COMET scores, comparing

the MT to the human references (Table 10): in both data settings, MTCUE matches the references to a significantly higher extent than the baseline.

5. Conclusions

We have argued for context-based personalisation of LMs by training a conditional generation architecture on dialogue accompanied by rich contextual annotations. We have also explored using such models to evaluate the context-specificity of professional and machine translations, providing insight into how well the generated translations are specific to the extra-textual context, without direct comparison to the human references. Finally, we have contributed CORNELL-RICH, a set of rich speaker and production annotations for a publicly available dialogue dataset. Below we summarise the findings specific to each research question.

RQA *How can rich character profiles be used to model the characters' speaking styles?*

The LMCUE architecture can be trained to exhibit personalisation leveraging the speaker and production metadata. This approach is effective in a few-shot scenario, matching speaker-specific fine-tuning performance, while requiring no such procedure. This makes it the more effective method if metadata is available.

RQB *How can a LM be personalised for a specific character solely by learning from data for characters with similar profiles?*

On a test set with unseen speakers, context-based personalisation yields a high SMRR score, (i.e. speaker separation). When both speaker and production metadata is used, the models achieve a reduction in perplexity comparable with the seen speaker scenario. Therefore, unlike speaker-specific fine-tuning, our method is applicable to new characters. We also found that having a varied pool of speakers and productions in training data correlates positively with SMRR.

RQB *Can MT offer personalisation benefits proportional to professional translations?*

Utilising speaker and metadata annotations in MT makes the language used in hypotheses more context-specific, as measured by the PMI score between such context and the generated text, when compared to a context-agnostic system. However, this context specificity is still stronger in gold standard (professional) translations. Our findings suggest that contextual LMs could be paired with automatic metrics for a more well-rounded evaluation of machine translation as they bring the aspect of the translations fitting the specific extra-textual context.

7.   Bibliographical References

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

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Dongyue Guo, Zichen Zhang, Peng Fan, Jianwei Zhang, and Bo Yang. 2021. A context-aware language model to improve the speech recognition in air traffic control. *Aerospace*, 8(11).

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2014. Enriching cold start personalized language model using social network information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 611–617, Baltimore, Maryland. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. 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.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

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.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. 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.

Trudy Milburn. 2004. Speech community: Reflections upon communication. *Annals of the International Communication Association*, 28(1):411–441.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2022. UserIdentifier: Implicit user representations for simple and effective personalized sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3449–3456, Seattle, United States. Association for Computational Linguistics.

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

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 253–260, New York, NY, USA. Association for Computing Machinery.

Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507(7490):9.

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

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium. Association for Computational Linguistics.

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.

Sebastian Vincent. 2023. *Context-Based Personalisation in Neural Machine Translation of Dialogue*. PhD thesis, University of Sheffield, Sheffield, UK.

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, Robert Flynn, and Carolina Scarton. 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*, pages 8210–8226, Toronto, Canada. Association for Computational Linguistics.

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. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages

1198–1212, Florence, Italy. Association for Computational Linguistics.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Wenhuan Zeng, Abulikemu Abuduweili, Lei Li, and Pengcheng Yang. 2019. Automatic generation of personalized comment based on user profile. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 229–235, Florence, Italy. Association for Computational Linguistics.

A. Details regarding the data collection campaign

The data collection process was carried out by two annotators, both native English speakers and experts in the dubbing and subtitling industry. After parsing CORNELL[5], a spreadsheet of characters was generated that included their name, source film, and the number of lines attributed.

From previous work (e.g. Johannsen et al., 2015) and hypotheses made based on experts' experience, we pre-defined a number of categories of information to collect about each character. Specifically, we selected categories that we hypothesised to (i) be identifiable from the available sources and (ii) influence a person's speaking style or vocabulary used. They were: their **age bracket** $\in$ {child, teen, young adult, adult, elderly}, **profession**, **character description** (a few sentences summarising their personality or character arc), **religion** and a **characteristic quote**: a typical or quotable thing the character might say. Additionally, the **gender** annotations from the original corpus were re-used, and an optional column "**additional information**" was included to collect comments

---

[5]https://convokit.cornell.edu/documentation/movie.html

from experts[6]. The characters with the most lines spoken were prioritised, resulting in all 714 characters with 100+ lines and 149 with 50−99 lines being annotated, totalling 863 characters.

Annotation sources Annotations are based on publicly available pages from Wikipedia[7] for individual films, as well as fan-made Fandom[8] pages for both films and characters. Where information was unavailable from these sources, the annotators either referred back to the corpus itself or skipped the given field altogether. The film metadata was obtained via the OMDb API.[9]

Annotation decisions The annotation process involved matching every script's name against an IMDb entry, which did not always yield a match as some scripts had been scrapped or rewritten or characters' names had been changed. Unidentifiable films and characters were not considered for annotation. Some information, especially *religion*, was occasionally difficult to find, in which case it would be skipped or labelled as *Unknown*. It was challenging to produce annotations for characters based on real people, or for a real person played by themselves. Where characters were based on historical figures, the annotators focused on the production interpretation of the person; when dealing with a characterisation of the person at a specific point in time, the focus was on their behaviour at that point in time. Finally, some characteristics were unsuitable for selected character information: e.g. when a character was immortal, it did not fit into set age brackets, and for some characters there were limited clues to determine their age bracket. In both cases, the final annotations were based on the annotators' expertise.

B. Details regarding the CORNELL-RICH Corpus

CORNELL-RICH comprises annotations for 863 speakers (speaker *profiles*), covering 135.7K utterances; nearly half of the annotated speakers have 150+ lines of dialogue and about 25% have 200+. At least 64.1% of conversational exchanges feature at least one annotated character and as much as 95.5% of the featured films are annotated with film metadata (Table 11). We provide a full list of collected metadata with examples in Table 12.

---

[6]Upon inspection: the annotators predominantly used this field to provide the actor's name, an interesting fact about the character (e.g. "Plays a caricature of himself"), or trivia.

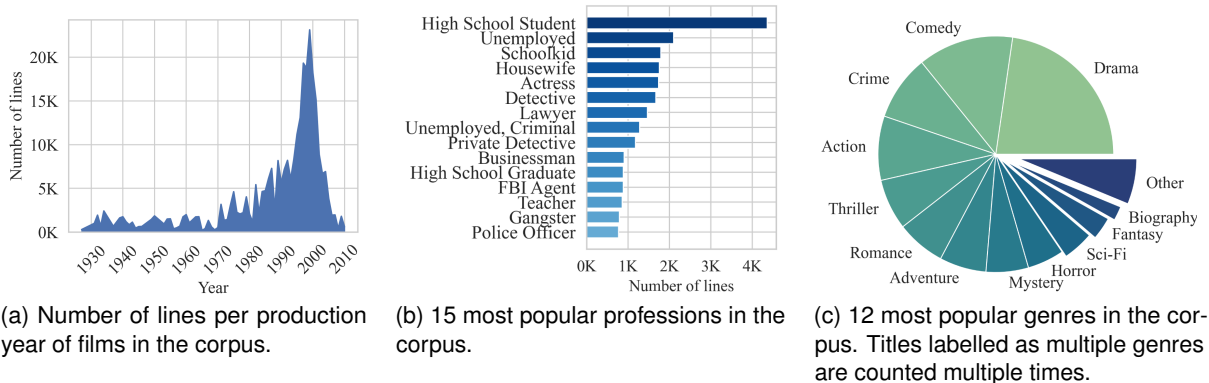[7]https://wikipedia.org/

[8]https://fandom.com/

[9]https://omdbapi.com/

(a) Number of lines per production year of films in the corpus.

(b) 15 most popular professions in the corpus.

(c) 12 most popular genres in the corpus. Titles labelled as multiple genres are counted multiple times.

Figure 4: Visualisation of a subset of features of the proposed corpus.

	Count	% of all	Utterances	% of all
speakers	9.0K	–	304.3K	–
👩🏾	863	9.5%	135.7K	44.6%
♟	8.2K	90.5%	170.1K	55.9%
exchanges	83.1K	–	304.3K	–
👩🏾 ⇔ (♟ or 👤)	53.3K	64.1%	202.4K	66.5%
♟ ⇔ 👩🏾	36.8K	44.3%	134.4K	44.2%
👩🏾 ⇔ 👩🏾	16.5K	19.8%	68.0K	22.3%
films	617	–	304.3K	–
annotated	589	95.5%	291.0K	95.6%

Table 11: Details of annotations compared to data quantities from CORNELL. 👩🏾 = speaker with rich annotations. ♟ = speaker without rich annotations.

Metadata type	Value
Speaker metadata	
Gender	A man
Age bracket	Adult
Profession	Attorney
Description	Galvin graduated from Boston College's law school. Galvin had a promising legal career ahead of him at an elite Boston law firm until he was framed for jury tampering by a partner due to his plans to expose the firm's underhanded activities. (...)
Quote	Your honor, with all due respect: if you're going to try my case for me, I wish you wouldn't lose it.
Country of origin	USA
Religion	Christian
Film metadata	
Genre	Comedy, Drama
PG Rating	PG Rating: R
Names of writers	Written by: Paul Andréota, André Cayatte, Henri Coupon
Country of production	France, Italy
Year of release	Released in 1974
Plot description	A French judge try to acquit a man who is accused of murdering his lover.

Table 12: A sample from CORNELL-RICH with each type of collected metadata.

As per Figure 4, the annotated films span nearly a century, with most lines coming from between the 1990s and 2005; the distribution of professions is significantly flatter, with the dominant field ("High School Student") only making up about 3% of the corpus. Finally, the most popular genres include drama, comedy, crime, and action.

C. Pre-training Strategy: Past Dialogue as Proxy for Metadata

This supplementary section presents empirical evidence that past dialogue can be used as a proxy for fine-tuning LMC_UE on speaker or production metadata. When fine-tuning, we use both speaker and production metadata. We report performance on test_unseen to also present sMRR scores.

	Pre-training		Fine-tune	PPL↓	sMRR↑
	✗ / ✓	Context			
BASE-LM	✓	–	✗	**28.78**	0.03
LMC_UE	✓	dialogue	✗	37.19	0.29
LMC_UE	✓	metadata	✗	30.95	**0.43**
BASE-LM	✗	–	✓	**39.60**	0.03
LMC_UE	✗	–	✓	51.14	0.03
BASE-LM	✓	–	✓	23.62	0.03
LMC_UE	✓	dialogue	✓	**22.31**	**0.96**
LMC_UE	✓	metadata	✓	22.71	0.89

Table 13: Results on test_unseen of CORNELL-RICH from different pre-training/fine-tuning setups. New results (top 5 rows) come from single runs.

As per Table 13, pre-training on OPENSUBTITLES leads to best perplexity when no context is used (BASE-LM), however using context yields improvements in sMRR, and these are stronger when metadata is used instead of dialogue. Similarly, without pre-training we also obtain the best perplexity with BASE-LM; here even sMRR remains at a baseline level, i.e. the contextual model fails to learn contextual dependencies correctly. Metadata only leads to superior results when both pre-training and fine-tuning are included

in the pipeline. Interestingly, pre-training on dialogue yielded the best results, though pre-training on metadata is not far behind ($+0.4$ PPL, -0.07 sMRR). We hypothesise that since past dialogue is much more diverse than film metadata (which contains many repeated fields), it is overall the better pre-training proxy for fine-tuning on new types of metadata, such as speaker profiles. For applications on other datasets, we therefore recommend pre-training on a similar dataset (domain-wise) with access to document-level information.

D.   Training details

D.1.   LMCUE and BASE-LM

We use off-the-shelf model architectures with predefined hyperparameters in FAIRSEQ and only tune on three values each for batch size (simulated 200K to 400K tokens) and learning rate ($3e-4$ to $1e-3$) based on validation performance on `valid` in CORNELL-RICH. For fine-tuning, we separately adapt these parameters for each dataset and metadata combination: learning rate ($5e-5$ to $1e-3$) and batch size (0.25K to 20K tokens). The best fine-tuning set of learning rate and batch size was $5e-5$ and 1.5K for LMCUE and $2e-3$ and 3K for BASE-LM. Each model was trained on a single 32GB V100 GPU with an early stopping condition of validation loss not improving for 5 epochs. Pre-training LMCUE and BASE-LM took 35 and 17.5 GPU hours respectively while fine-tuning these models took respectively 0.78 and 0.32 GPU hours on average.

D.2.   MTCUE and BASE-NMT

The machine translation models were trained in a similar way to the language models. We use the dataset and follow the pre-training regimen described in Vincent et al. (2023). For fine-tuning, we separately adapt these parameters for each dataset and metadata combination: learning rate ($5e-5$ to $1e-3$) and batch size (1K to 80K tokens). The best fine-tuning set of learning rate and batch size was $1e-3$ and 80K for MTCUE and $3e-3$ and 20K for BASE-NMT. Each model was trained on a single 32GB V100 GPU with an early stopping condition of validation BLEU not improving for 5 epochs.

E.   Cost-benefit Analysis of Human Annotations

Granular manual annotations are costly to obtain. Cost-benefit analysis helps avoid the misallocation of limited annotation funding and resources. This section presents the results of the cost-benefit analysis we conducted to show which individual speaker attributes produce the most **benefit** (reduction in perplexity) w.r.t. the perceived **cost** of producing them.

We asked the two human annotators to assess the effort required for the annotation task using three metrics on a Likert scale of 1 to 10: *access* (how difficult it was to find information), *credibility* (how confident they were in the accuracy/usefulness of the information), and *time* (how much time was needed relative to other fields). We took the mean of both annotators' scores after reversing credibility ($C = 10 - C + 1$). We then conducted a simple experiment to measure the **benefit** of each metadata type by fine-tuning the pre-trained LM-CUE on each speaker metadata type evaluated individually. Finally, we measured the reduction in perplexity from including this information (Figure 5) compared to the 91M parameter decoder in LMCUE, since that is the decoder we are trying to improve with context.

The figure suggests that *description*, *profession* and *quote* yield the greatest perplexity reduction in both datasets, around 5 to 6%. *Description*, the best-performing attribute, alone achieves $88.7/91.9\%$ of the perplexity reduction of LMCUE ($\mathcal{S}$). On the other hand, *age bracket*, *religion* and *country of origin* yield the smallest improvements, and a better improvement can be achieved with the parameter-matched BASE-LM. For CORNELL-RICH, they still help marginally (1 to 2%), while for ZOO improvements from *age* and *religion* are negligible. This analysis suggests why King and Cook (2020), who implemented context-based adaptation using only *age* and *gender*, found it inferior to other methods; we found other variables such as *description* to be significantly more useful.

Other findings of the analysis are consistent among the two corpora. *Profession*, collected at a relatively small cost, is one of the top-3 attributes for both, hence the most cost-effective. *Religion* is the least cost-effective attribute, requiring the most effort but producing the least benefit. Finally, experimental attributes such as characteristic *quote*s and *additional information*[10] have been shown to be useful, the latter placing in the middle of the ranking whilst the former is on par with the best attribute for CORNELL-RICH.

F.   Example from the ZOO-MULTI corpus

In Table 14 we show what a sample from the ZOO-MULTI corpus looks like, including an example for each annotation as well as how many samples

---

[10]Since *additional information* was not collected for ZOO, it is not present on the ZOO plot of Figure 5.
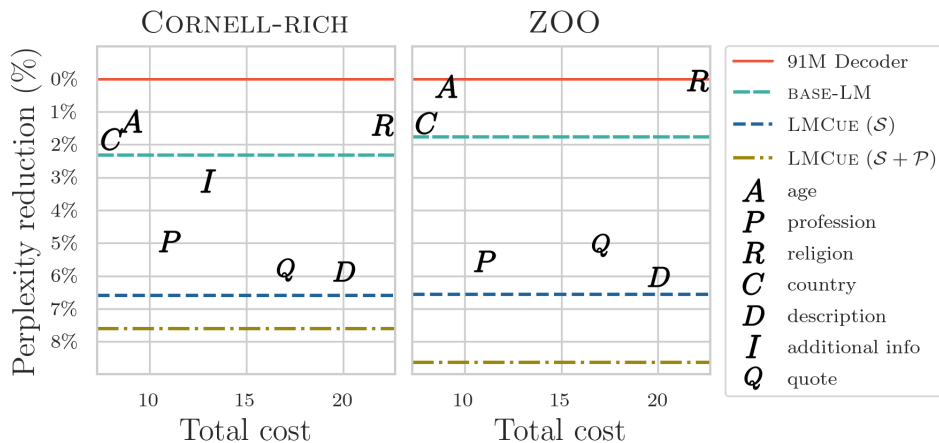
Figure 5: Perplexity reduction from training LMCUE with individual speaker attributes.

were annotated with the given context type.

Type of text field	Anno. %	Example
English (source)	–	That will take 12 hours.
French translation	–	Cela prendra 12 heures.
German translation	–	Es wird 12 Studen dauern.
Polish translation	–	To zajmie 12 godzin.
Production		
country	87.1%	United States
genre	87.1%	Drama, History
PG rating	36.4%	PG rating: TV-14
plot description	75.9%	U.S. fighter pilots are recruited to test experimental aircraft and rockets to become first Mercury astronauts.
writers' names	31.6%	Written by: Mark Lafferty
year	87.1%	Released in 2020
Speaker		
age bracket	50.3%	Adult
description	50.3%	Chris Kraft is an American aerospace and NASA engineer who was instrumental in establishing the agency's Mission Control Center and shaping its organization and culture. He is the Flight Director in charge of getting the mission control team ready for the launch of the 'Mercury Seven' team.
characteristic quote	39.5%	He just kited a damn check.
country of origin	50.3%	United States
gender	95.5%	I am a man
profession	50.3%	Flight Director
religion	13.5%	Christian

Table 14: Details regarding the ZOO-MULTI corpus. Quantity refers to the number of samples within the training sets. Certain examples have been shortened for brevity.

G. Examples of Evaluation

In this section, we show some examples where the hypotheses from the contextual model (MTCUE) were deemed more contextual (i.e. higher PMI) than the hypotheses from BASE-NMT. All outputs come from systems in the OVERLAP setting and the English-to-Polish translation direction.

Example A

Context Source	*Family cooking competition show.* We're okay, we're doing just fine, just...
Reference	Wszystko jest w porządku. *('Everything is alright.')*
BASE-NMT	(✗) Nic nam nie jest. *('We are fine.')* -0.35 PMI -1.29 COMET
MTCUE	(✓) Radzimy sobie. *('We're coping.')* **0.32** PMI $-$**0.10** COMET

Example A exhibits a behavioural agreement adaptation. During cooking show walkthroughs there is frequently talk of managing stress and working under pressure. It is done in good faith to let the contestants discuss their struggle. In Example 1, the non-contextual translation *Nic nam nie jest.* would be far less likely to occur in this walkthrough setting than *Radzimy sobie.* as this phrase is applicable to immediate impact situations (e.g. falls) rather than stressful, ongoing situations like a competition show. It also has a more defensive tone which is not a good fit for a lighthearthed family series.

Example B

Context Source	*Family cooking competition show.* Best think up some new recipes.
Reference	Potrzebujemy nowych przepisów. *('We need new recipes.')*
BASE-NMT	(✗) Lepiej wymyśl nowe przepisy. *('[You$_s$] better think up some new recipes.')* -0.76 PMI **0.59** COMET
MTCUE	(✓) Wymyślmy nowe przepisy. *('Let's think up some new recipes.')* $-$**0.51** PMI 0.48 COMET

In **Example B**, MTCUE guesses the number of addressees correctly: in this family cooking com-

petition show, when someone is congratulated it is usually a family, i.e. a group of people. In our case-by-case analysis we continually found that the personalised LM exhibits a preference for correct gender and plurality.

Example C

Context	*Family cooking competition show.*
Source	Try the balls.
Reference	Spróbuj kulkę. *('Try a ball.')*
BASE-NMT	(✗) Spróbuj piłeczek. *('Try the footballs.')*
	-0.55 PMI **0.51** COMET
MTCUE	(✗) Spróbuj jajek. *('Try the eggs.')*
	1.02 PMI -0.19 COMET

Finally, in **Example C**, the language model assigned a significantly stronger PMI score to the hypothesis which translated *balls* as *eggs* instead of *footballs*. While neither hypothesis is correct (the "balls" in the original text most likely referred to meatballs, or dough balls), the word *eggs* is highly specific to the context of cooking, and therefore rated very highly by the personalised LM. This example shows that our method must be used in tandem with a standard translation quality metric such as BLEU or COMET, as it is a monolingual evaluation tool that disregards the faithfulness to the source sentence.